Dean&Francis
ISSN 2959-6157

Construction and Analysis of a Risk Prediction Model for Heart Disease Based on Binary Logit Regression and Random Forest Model

Ailin Du

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
Corresponding author: ailin.du24@student.xjtlu.edu.cn

Abstract:

Heart disease (HD) is one of the most serious health problems worldwide. If it can be predicted at an early stage, it can prevent heart attacks and thereby reduce the mortality rate of HD. Previous studies that successfully predicted HD using random forest models and binary logistic models have inspired this research. Therefore, this research jointly uses the random forest model and the binary logistic model to make more reliable and stable predictions of HD. This research analyzed the survey data on the annual health status of over 400,000 adults from the Centers for Disease Prevention and Control of the United States in the public database of the Kaggle website in 2022. Finally, it was concluded that stroke is the most significant factor affecting HD. In the visual analysis and the predictions of the two models, the risk factor Stroke stood out significantly. In the prediction results of the random forest model, this factor ranked among the top four, with a model prediction accuracy of 89%. In the binary logit model prediction, the variable Stroke ranked first, with a model prediction accuracy of 91.36%. Both models had relatively high accuracy rates, and Stroke was determined as a significant influencing factor in both models, making the prediction results reliable. This research provides a theoretical basis for early clinical screening of HD and offers more innovative and reliable prediction methods, which is expected to reduce the mortality rate of HD in the future.

Keywords: Random Forest Model, Binary Logit Regression, Cardiovascular prediction, Heart Disease Prediction

ISSN 2959-6157

1. Introduction

Cardiovascular diseases (CVDs) have received widespread attention worldwide. The diseases are the leading cause of death globally and one of the most serious health issues. The consequences of these diseases in terms of death and disability are severe, and they also impose a burden on the economy. [1] Globally, in 2016, CVDs ranked among the top five causes of total expected life loss years, and CVDs accounted for nearly one-third of all deaths [2]. Approximately seven million people die from CVDs each year, which causes a significant economic burden globally [3].

Early detection of heart disease (HD) can prevent fatal events such as heart attacks. Machine learning models have a significant impact in heart disease prediction, enabling the early detection of heart diseases and timely intervention. The simple algorithm of logistic regression also performs well in the early prediction of heart diseases [4]. Therefore, many scholars have conducted relevant research on the early prediction of HD. Kellen Sumwiza et al. based on the random forest model, proposed an effective integration method by combining multiple feature selection techniques to improve accuracy. Their model has an accuracy rate of up to 99%, which represents a significant improvement compared to other models [5].

Ebtehag Mustafa Mohammed et al. analyzed the data obtained from the Madani Heart Center in Sudan in 2019 and compared the binary logic model with the neural network model to determine the risk factors for CVDs. This study concludes that although the prediction accuracy of the artificial neural network is higher than that of the binary logistic regression model, the two methods have similar significance and importance in terms of the influence of the independent variables in the analysis [6].

Given the successful application of the binary logit regression model and the random forest model in the field of HD prediction, this study innovatively adopted a multi-model consensus strategy, that is, applying these two models independently, systematically comparing and extracting the common points in the analysis results of the two models, and more robustly identifying the factors that significantly affect HD. This study aims to use this more innovative and robust approach to identify high-risk factors associated with HD, thereby achieving early prediction of HD.

2. Methods

2.1 Data Source

The data for this study were obtained from Kaggle. It is based on the annual health status survey data of over 400,000 adults conducted by the Centers for Disease Control and Prevention of the United States in 2022. This data was collected by the Centers for Disease Control and Prevention through telephone surveys, and this system is the largest ongoing health survey system in the world [7]. The dataset contains 18 variables. This study selected 49,947 samples and 13 variables from this dataset for research. The binary variable HD was set as the dependent variable, and the remaining 12 variables were considered as factors that might affect the dependent variable.

2.2 Indicator Selection and Explanation

The 13 variables used in this study, along with their corresponding explanations, are presented in Table 1.

Table 1. Variables and their explanations

Variables	Explanations
Heart Disease	Whether one has heart disease, 1=Yes, 0=No
BMI	Body mass index
Smoking	Whether one smokes, 1=Yes, 0=No
Stroke	Whether one has stroke, 1=Yes, 0=No
Diff Walking	Diff Walking is the abbreviation of Difficult Walking. Whether difficult to walk, 1=Yes, 0=No
Sex	1=Male, 0=Female
Age Category	Different numbers represent different age groups. 1=18-24, 2=25-29, 3=30-34, 4=35-39, 5=40-44, 6=45-49, 7=50-54, 8=55-59, 9=60-64, 10=65-69, 11=70-74, 12=75-79,13= 80 and above
Diabetic	Whether one has diabetes, 1=Yes, 0=No
Physical Activity	Whether one has physical activities, 1=Yes, 0=No
Sleep Time	The sleep duration of the survey participants (recorded in hours)
Asthma	Whether one has asthma, 1=Yes, 0=No
Kidney Disease	Whether one has kidney disease, 1=Yes, 0=No
Skin Cancer	Whether one has skin cancer, 1=Yes, 0=No

In the original dataset, the statistical values under the Diabetic variable included not only the regular data on whether the subject had diabetes, but also the data on whether they had diabetes under the conditions of bor-

derline diabetes and during pregnancy. However, since these types of data accounted for a very small proportion of the total data, this study excluded these two conditions and directly recorded whether the research subjects had diabetes, thereby converting the data under the Diabetic variable into binary classification data, which is easier to handle. The data under the Age Category variable was transformed from age groups into continuous data for convenient analysis.

2.3 Method Introduction

This study employed two models, namely the binary logit regression model and the random forest model. The binary logit regression model analysis can explain the influence of independent variables on the dependent variable through results such as P values and OR values, thereby identifying the key variables that affect HD. The OR values in the analysis results can directly quantify the risk. By observing the 95% confidence interval of OR, if the interval does not include 1, it is considered a significant influencing factor. The formula of the binary logit model is:

$$ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k. \quad (1)$$

Where P represents the probability that the dependent variable y equals 1. β_0 represents the intercept term, while $\beta_1,\beta_2,\ldots,\beta_k$ are the regression coefficients corresponding to the respctive variables $x_1,x_2,\ldots,x_k$.

The random forest model can determine the factors that significantly affect HD by analyzing the feature weights in the results. The sample ratio of the training set to the test set of the model is eight to two. The prediction effect of the model is judged by analyzing the precision, recall rate and F1-score in the results.

3. Results and Discussion

3.1 Visualized Analysis

Before conducting the systematic analysis, to preliminarily explore the potential risk factors of heart disease, some multi-dimensional visual analyses were first performed on the data. For continuous variables, scatter plots were used for intuitive presentation, as shown in Fig. 1.

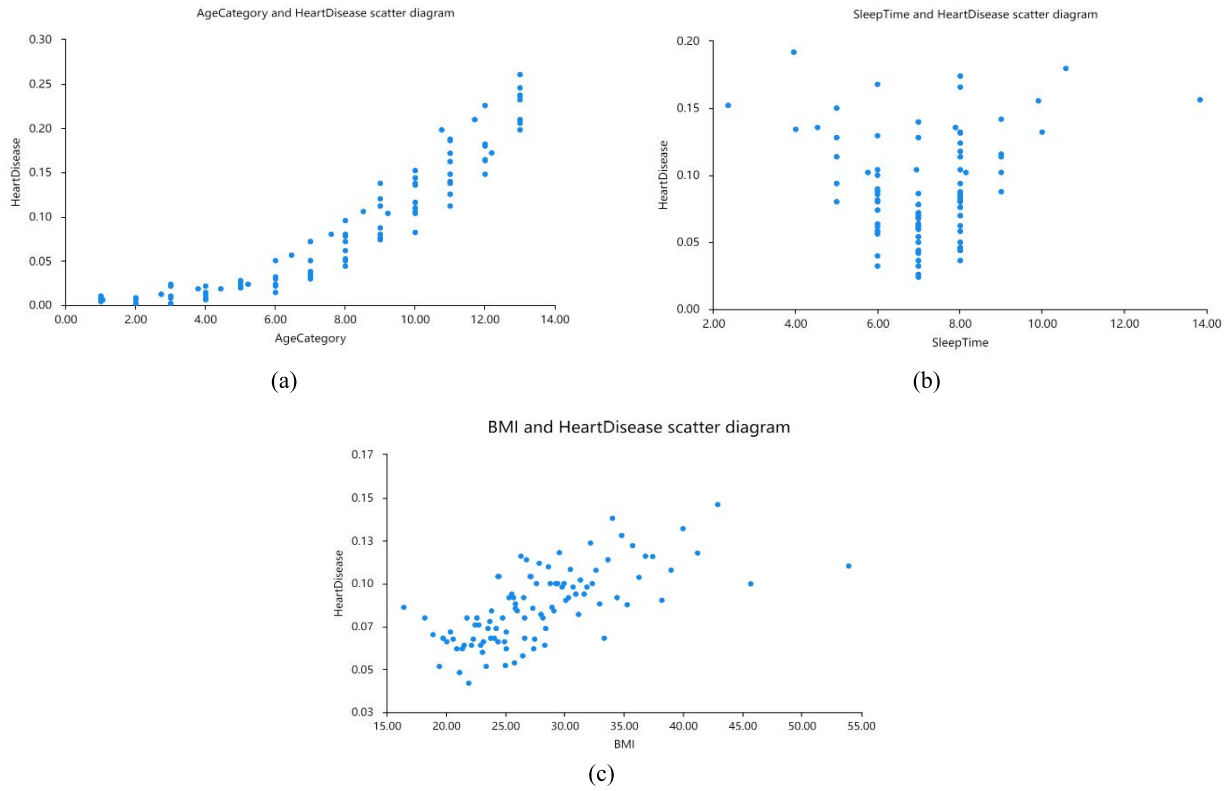

Fig. 1 Scatter plot of HD versus three continuous independent variables, (a): Age Category and HD scatter diagram; (b): Sleep Time and HD scatter diagram; (c): BMI and HD scatter diagram (Photo/Picture credit: Original).

ISSN 2959-6157

From Fig. 1, it can be clearly seen that age is strongly positively correlated with the risk of HD. As age increases, the risk of HD continues to rise. This suggests that age is a risk factor for HD. BMI does not show a simple linear relationship with whether one has HD, but overall, the area with a high BMI has a higher risk of HD. This indicates that obesity remains a risk factor for HD. The scatter plot of sleep shows a clear U-shaped pattern. The lowest point is approximately at a sleep duration of seven hours. Shorter or longer sleep times will have a higher risk of HD. A sleep duration of around seven hours is relatively optimal and can minimize the risk of HD. Then, a clustered plot was drawn for the binary categorical variables in the data for analysis, as shown in Fig. 2.

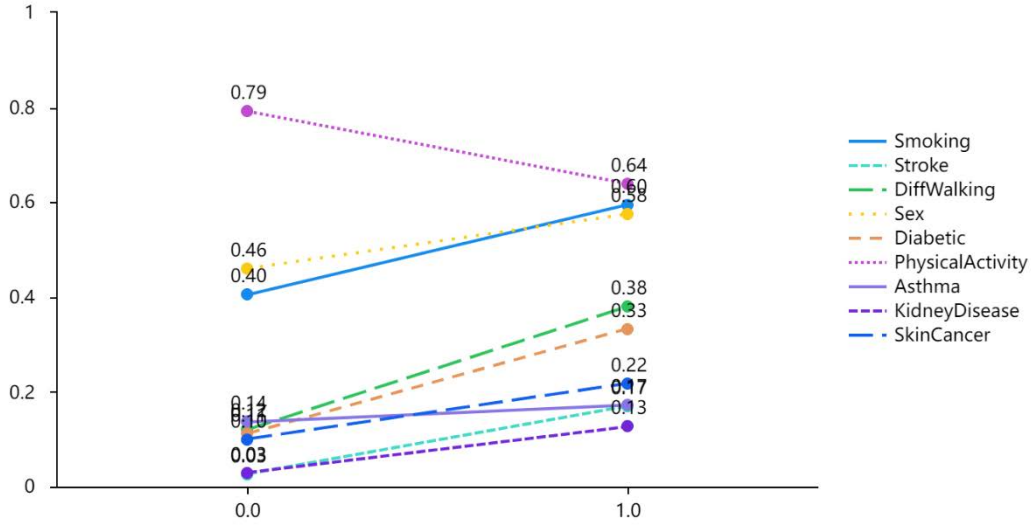

Fig. 2 HD and comparative analysis of all Binary categorical variables (Photo/Picture credit: Original).

From Fig. 2, it can be observed that the relatively steep lines may be associated with whether one has HD, such as Physical Activity, Smoking, Stroke, and Diabetes. Among them, Physical Activity shows a negative correlation with having HD, while the other three variables show a positive correlation. In order to further intuitively explore which variables are related to HD, the final part of this study conducted a Spearman correlation analysis, as shown in Fig. 3.

Fig. 3 Spearman correlation analysis heatmap (Photo/Picture credit: Original).

By observing the color intensity of each variable in the heat map, this study can initially identify that the variables such as Age Category, Stroke, Diff Walking, and Diabetic have a strong correlation with whether one has HD.

Through the above-mentioned several visual analyses, this study have a general understanding of the influencing factors of HD. However, in the future, a specific and systematic analysis of this study is needed to find the answers.

3.2 Binary Logit Regression

The above-mentioned visualization results indicate that there is a correlation between the risk of HD and various factors. To quantify the direction and intensity of these effects, this study employed a binary logit regression model for analysis, and the resulting analysis data are presented in Table 2.

Table 2. Summary of binary logit analysis results

Items	Regression coefficient	SE	z	Wald χ^2	p	OR	OR 95% CI
BMI	0.012	0.003	4.316	18.630	0.000	1.012	1.007 ~ 1.018
Smoking	0.432	0.035	12.245	149.932	0.000	1.540	1.437 ~ 1.651
Stroke	1.287	0.055	23.412	548.101	0.000	3.623	3.253 ~ 4.036
DiffWalking	0.677	0.041	16.324	266.458	0.000	1.967	1.814 ~ 2.134
Sex	0.688	0.036	19.195	368.453	0.000	1.990	1.855 ~ 2.135
AgeCategory	0.270	0.007	36.696	1346.633	0.000	1.310	1.291 ~ 1.329
Diabetic	0.715	0.041	17.621	310.500	0.000	2.044	1.888 ~ 2.213
PhysicalActivity	-0.169	0.040	-4.275	18.278	0.000	0.844	0.781 ~ 0.912
SleepTime	-0.046	0.011	-4.270	18.231	0.000	0.955	0.935 ~ 0.975
Asthma	0.262	0.048	5.432	29.511	0.000	1.300	1.182 ~ 1.429
KidneyDisease	0.715	0.060	11.847	140.350	0.000	2.043	1.816 ~ 2.300
SkinCancer	0.176	0.045	3.905	15.249	0.000	1.192	1.092 ~ 1.302
intercept	-5.805	0.143	-40.458	1636.849	0.000	0.003	0.002 ~ 0.004
McFadden $R^2 = 0.201$							

To identify variables significantly impacting HD, P value, OR value, and regression coefficient must be considered together. A variable is significant if its P value < 0.01 (non-significant if P value > 0.05) or its OR's 95% confidence interval does not cross 1; an OR > 1 means a higher HD risk with the independent variable's increase (and vice versa for OR < 1), with a larger OR indicating a stronger effect. A positive regression coefficient signifies higher HD risk as the independent variable rises (negative for lower risk), and a larger absolute coefficient means greater influence.Through comprehensive analysis of the data, it can be concluded that, Stroke, Diabetes, Kidney Disease and Sex are the top four significant influencing factors. Among them, Stroke ranks first as the risk factor. After ischemic stroke, cardiovascular complications face a serious burden. In the days after the stroke, approximately 10% to 20% of patients will experience adverse heart problems [8]. Among many young patients with acute stroke, the prevalence of traditional CD risk factors that coexist in them has been on the rise [9].

Furthermore, it was noted that the regression coefficient of the variable Sleep Time in the binary logit model's analysis results was negative. This is inconsistent with the results of this research's previous scatter plot and correlation analysis. The scatter plot indicated that both too short or too long sleep time could lead to an increase in risk. Simple linear correlation analysis might not be able to capture this complex pattern. Therefore, the positive correlation it obtained might only be a partial reflection of multiple mixed factors. However, the binary logit regression model, after controlling for various confounding factors, concluded that there is a negative correlation between sleep duration and the risk of illness. This conclusion might be more reliable and can better eliminate misleading factors.

The McFadden R^2 value of the model is 0.201, greater than 0.2. Its value is usually considered to be within the range of 0.2 to 0.4, which is generally regarded as a well-fitted model. In the likelihood ratio test results of the binary Logit regression model, the P-value of the model is 0, which is less than 0.05. This indicates that the model is valid. The overall prediction accuracy of the model is 91.36%, and the model effect is acceptable.

3.3 Random Forest Model

In order to further verify the prediction results obtained from the binary logit regression model and to obtain more

ISSN 2959-6157

reliable conclusions, this study employed the random forest model, together with the binary logit model, to jointly identify the risk factors. In the analysis of the random forest model, the feature weight map obtained is shown in Fig 4.

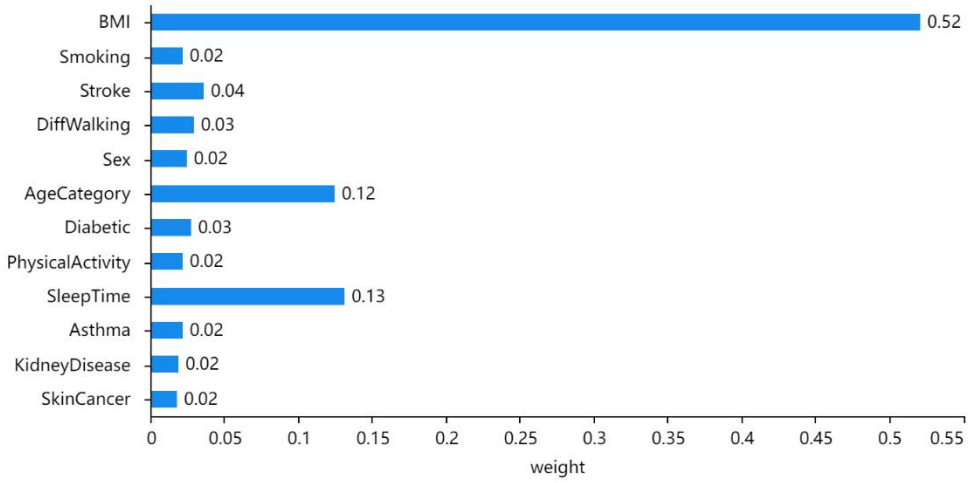

Fig. 4 Feature weight map (Photo/Picture credit: Original).

From Fig. 4, it can be clearly seen that the top four influencing factors are BMI, Sleep Time, Age Category, and Stroke. The data regarding the model's performance is shown in Table 3.

Table 3. Evaluation results of the test set model

Items	Accurate Rate	Recall Rate	f1-score	Sample size
0.0	0.92	0.96	0.94	9110
1.0	0.25	0.12	0.17	880
Accurate rate			0.89	9990
Mean	0.59	0.54	0.55	9990
Mean(synthesize)	0.86	0.89	0.87	9990

F1-score is a comprehensive evaluation metric that combines precision and recall rates. Generally, the closer this value is to 1, the better the model performance. In this study, the overall prediction accuracy of the random forest model was 89%, and the model performance was acceptable. In this model, the prediction accuracy for the sample of having heart disease (i.e., 1.0 in the table) is relatively low. This phenomenon is mainly due to the significant difference in the sample sizes between those with heart disease and those without. The imbalance in sample size categories is the cause of this.

4. Conclusion

Based on the analysis of the binary logit regression model and the random forest model, it was found that the variable Stroke was jointly determined by both models as a high-risk factor for HD. Moreover, this variable also performed well in the previous correlation analysis and cluster diagram. Therefore, this study concludes that Stroke is a significant factor influencing HD. However, this study still has some limitations. It is noted that the prediction accuracy of both models for the samples without HD during the prediction process is not very high. This might be because the number of patients with and without HD in the dataset is unbalanced for this variable. However, the overall prediction accuracy of both models is relatively high, and the analysis effect is good. It has been predicted that Stroke is a risk factor for HD. This indicates a close association between stroke and HD. This study provides a theoretical basis for the early screening of high-risk groups for HD, offers key indicators for screening, making the screening more efficient, and also provides a basis for early clinical diagnosis, saving costs.

References

[1] Amini M., Zayeri F., Salehi M. Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study

2017. BMC Public Health, 2021, 21(1): 401.
[2] Bourne R. R., GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. 2023.
[3] Ralapanawa U., Sivakanesan R. Epidemiology and the magnitude of coronary artery disease and acute coronary syndrome: a narrative review. Journal of Epidemiology and Global Health, 2021, 11(2): 169-177.
[4] Bhowmik P. K., Miah M. N. I., Uddin M. K., Sizan M. M. H., Pant L., Islam M. R., Gurung N. Advancing heart disease prediction through machine learning: techniques and insights for improved cardiovascular health. British Journal of Nursing Studies, 2024, 4(2): 35-50.
[5] Sumwiza K., Twizere C., Rushingabigwi G., Bakunzibake P., Bamurigire P. Enhanced cardiovascular disease prediction model using random forest algorithm. Informatics in Medicine Unlocked, 2023, 41: 101316.
[6] Mohammed E. M., Osman E. G. A. Comparison between neural networks and binary logistic regression for classification observation (case study: risk factors for cardiovascular disease). Proc. 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), IEEE, 2021: 1-6.
[7] Pytlak K. Indicators of heart disease (2022 update. Kaggle, 2022. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease (Accessed: 2025-07-22).
[8] Scheitz J. F., Sposato L. A., Schulz-Menger J., Nolte C. H., Backs J., Endres M. Stroke–heart syndrome: recent advances and challenges. Journal of the American Heart Association, 2022, 11(17): e026528.
[9] George M. G. Risk factors for ischemic stroke in younger adults: a focused update. Stroke, 2020, 51(3): 729-735.