Visualization analysis and logistic regression based heart disease risk prediction

Haoyu Zhao

Department of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, Shandong, 250014, China

631402070305@mails.cqjtu.edu.cn

Abstract. Heart-related illness is the major cause of global death. The optimal solution to tackle this problem and to improve public health is early detection and prevention. Manually diagnosis is tedious and time consuming, which is difficult to be applied for large scale medical inspections, and hence machine learning, computer-based automatic algorithms could be adopted. Logistic regression is a commonly used statistical method for predicting the risk of binary outcomes, such as the presence or absence of heart problems. In this study, logistic regression is leveraged to a dataset of medical records. It is not only developed as an effective model for the early detection of heart disease, but also leveraged for identifying the crucial risk factors of the disease. The results showed that the logistic regression model achieved a high level of accuracy for heart risk prediction, which overall accuracy is 85%. Factors including sex, cholesterol level, age, and blood pressure are observed possessing highest correlations with heart disease.

Keywords: machine learning, heart disease, logistic regression.

1. Introduction

Heart failure is a medical condition. Clinically, it is usually presented as the insufficiently pumped blood by heart, which could potentially arouse symptoms like fatigue and shortness of breath [1,2]. It is a serious medical condition affecting millions of people globally. There are many underlying conditions including heart failure, high blood pressure, and heart valve problems. It is a major cause of hospitalization and can even lead to sudden cardiac death if left untreated [3,4].

Traditionally, heart failure diagnosis and management have relied heavily on manual assessments and clinical judgment. However, recent advances in machine learning offer new opportunities to improve the accuracy and efficiency of diagnosis and treatment [5,6]. Through the use of machine learning, vast amounts of data can be processed and patterns can be identified that may not be easily discernible by human analysis. This can help clinicians make more accurate diagnoses and develop more personalized treatment plans [7,8].

In this study, the usage of logistic regression as a machine learning algorithm is validated for predicting heart failure. By using this algorithm, the paper aims to improve the accuracy of heart failure prediction and provide a more efficient means of diagnosis and management. Furthermore, by leveraging machine learning, it is expected to identify previously unknown risk factors and biomarkers that may be used in future research and clinical practice.

2. Method

This work selected a dataset of medical records from a hospital, including information on patients' age, gender, blood pressure, cholesterol level, smoking status, and other clinical variables. The correlations of between each factor pairs and the classification of disease is implemented by logistic regression algorithm. This work used a stepwise selection method to find out the most important risk factors related to the development of cardiovascular disease. Moreover, a logistic regression model is implemented to predict the risk of heart disease.

2.1. Dataset

The dataset is achieved from the UCI open-source dataset [9]. The dataset consists of 918 messages and includes 12 features: 6 integer, 1 floating-point, and 5 object data types. Notably, all data is complete and intact. The dataset contains information on individuals who have or have not experienced heart failure, with features including age, sex, blood pressure, serum cholesterol, and chest pain type. Before analyzing the data, the dataset was examined to gain insight into its characteristics, including the distribution of the features and the prevalence of heart failure.

To begin the analysis, the commonly used libraries for data analysis were imported, including NumPy, Pandas, and Scikit-Learn. The dataset was imported and its characteristics were examined. The dataset consists of X observations and Y variables, including the target variable indicating the presence or absence of heart failure. Before applying any machine learning algorithms, the dataset was preprocessed to handle missing values and outliers, and feature scaling was performed to standardize the variables. Specifically, there are missing values that hinder the application algorithms. During pre-processing they were imputed as the mean or median of the corresponding feature, depending on the data type. The work detected and removed outliers using the Z-score method. Besides, features are rescaled with standard scaler, which rescales the features to zero mean and unique variance, which allows all features are at comparable scale. This preprocessing step regularized data format for machine learning algorithms.

2.2. Model

In this study, logistic regression is employed as the classification model for predicting heart failure. Logistic regression is a widely used linear model for binary classification tasks, which predicts the probability of a binary outcome (in this case, the occurrence of heart failure) based on the input features. It is a popular choice due to its simplicity and interpretability, as well as its ability to handle both categorical and continuous input features [10].

The mathematical principle of logistic regression is to model the logarithm of the odds ratio (logit) as a linear combination of the input features. An odds ratio is a measure of the likelihood of an event occurring compared to the likelihood of the event not occurring. By modeling the logit as a linear combination of the input features, logistic regression can estimate the probability of an event occurring given the input features.The values of the model parameters are computed by maximizing the likelihood of observing the training data through the method of maximum likelihood estimation.

One advantage of logistic regression is that it is a simple and interpretable algorithm with feature coefficients representing the degree of influence of the corresponding feature on the outcome. Logistic regression can also handle both categorical and continuous input features. However, logistic regression also has some limitations. It is assumed that there exists a linear relationship between the input features and the logit; however, this assumption may not always hold in practical situations. Additionally, logistic regression may not perform well when there are many input features or when the input features are highly correlated.

In the context of logistic regression, C and penalty are two important hyperparameters that can influence the outcome of the model. Hyperparameter C controls the inverse of the regularization strength, with smaller values of C indicating stronger regularization. Penalty refers to the type of regularization used, with L1 and L2 regularization being the most commonly used types. L1 regularization promotes sparsity in the coefficients of the model, while L2 regularization encourages small coefficients. The hyperparameters were tuned using grid search cross-validation, which involves evaluating the model's performance on different hyperparameter combinations that performs the best during validation.

In the study, several metrics were utilized for performance evaluation, including precision, recall, F1-score, and accuracy. The accuracy metric measures the overall correctness of the model's predictions. Precision calculates the fraction of correctly predicted positive samples over the entire positive predictions. Recall measures the fraction of true positives among all actual positives. The F1-score is the harmonic mean of precision and recall, which provides a balance between the two metrics, calculated as their harmonic mean. These performance measures will provide a holistic view of the model's ability to predict heart failure, and they will be utilized to compare the effectiveness of various models in the study.

To train the logistic regression model, the dataset was divided as training and testing sets with the ratio at 3 v.s.1. To train it, the training samples were used, while the testing samples were employed to assess performances. The logistic regression model was programmed utilizing the scikit-learn library in Python, and grid search with 5-fold cross-validation was performed to tune the hyperparameters C and penalty on the training set. The optimal hyperparameters were chosen by selecting the highest F1-score from the validation set.

After tuning the hyperparameters, the logistic regression model was retrained using the entire training set with the best hyperparameters. The model's performance was assessed on the testing samples leveraging the evaluation metrics mentioned earlier. In addition, a confusion matrix was also generated to deliver deeper understandings of outcomes.

3. Result

3.1. Visualization analysis

The visualization analysis is an effective method to explore the relationships between different variables. In this study, a heatmap was constructed to visualize the correlation between different variables using python, illustrating in Figure 1. Highly correlated variables are represented by a strong color gradient on the heatmap.

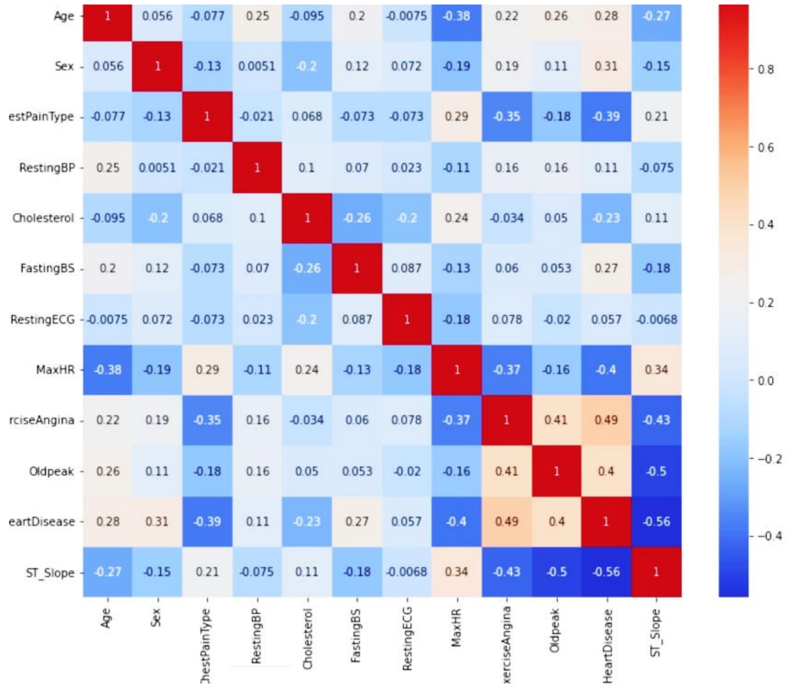

Figure 1. Heatmap of feature correlation.

From the heatmap, it can be observed that there is a stronger correlation between heart failure and the characteristics of ST_Slope, ExerciseAngina, MaxHR, and Oldpeak. Additionally, a certain correlation is observed between the characteristics of Age, Sex, and RestingBP.

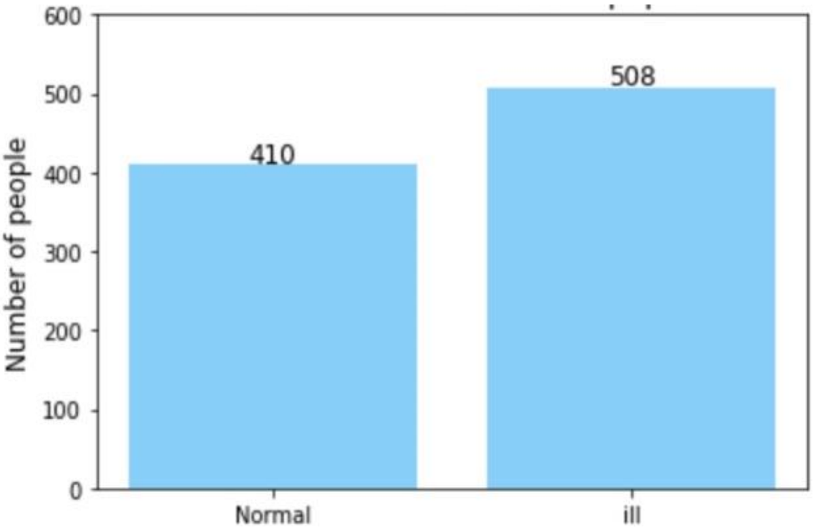

Figure 2. Distribution of sick and normal population.

According to the visualization analysis in Figure 2, there are 508 patients and 410 non-patients in the dataset. To represent this result, a tree chart is leveraged to visualize the distribution of patients and non-patients in the dataset.

Examining the relationship between heart disease and age, the purpose of Figure 3 is to segment the ages of patients with heart disease and merge the segmentation results with information on whether each patient has heart failure, to perform predictive analysis on heart failure. The results reveal that middle-aged age group dominates the dataset, indicating that middle-aged people are more likely to be susceptible to heart disease.

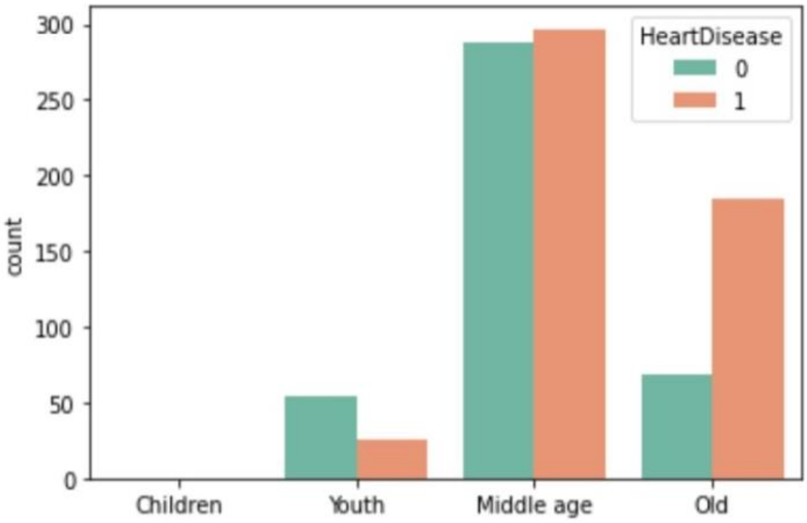

Figure 3. Relationship between heart disease and age.

The relationship between heart disease and sex was also examined through the visualization analysis in Figure 4. The results indicated that the prevalence rate of heart disease was higher among women than men.

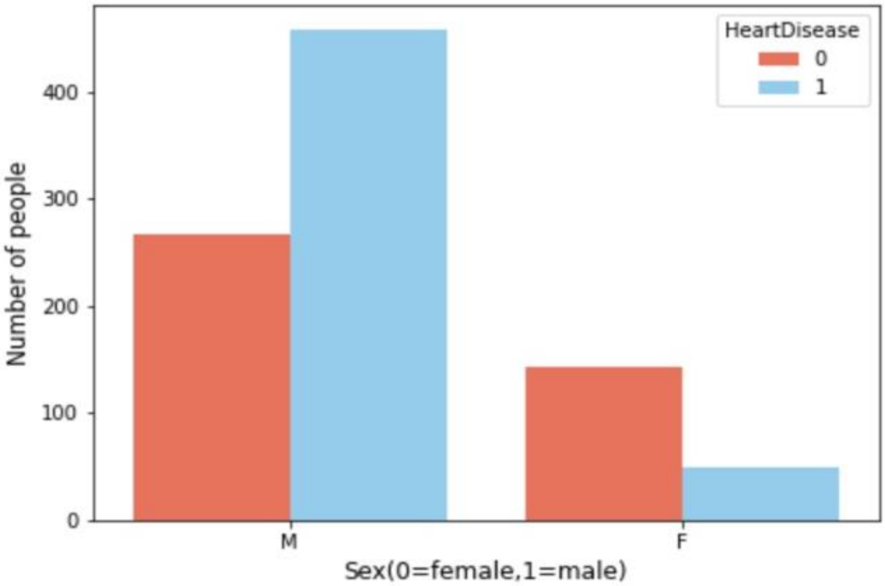

Figure 4. Disease distribution of different genders.

The analysis reveals that there is a relationship between the incidence of heart disease and maximum heart rate during exercise, as illustrated in Figure 5. Specifically, the probability of being diagnosed with heart disease is relatively high for individuals who experience angina pectoris caused by exercise. Therefore, individuals with a history of heart disease should avoid engaging in excessive physical activity. The findings are presented in the form of a tree diagram.

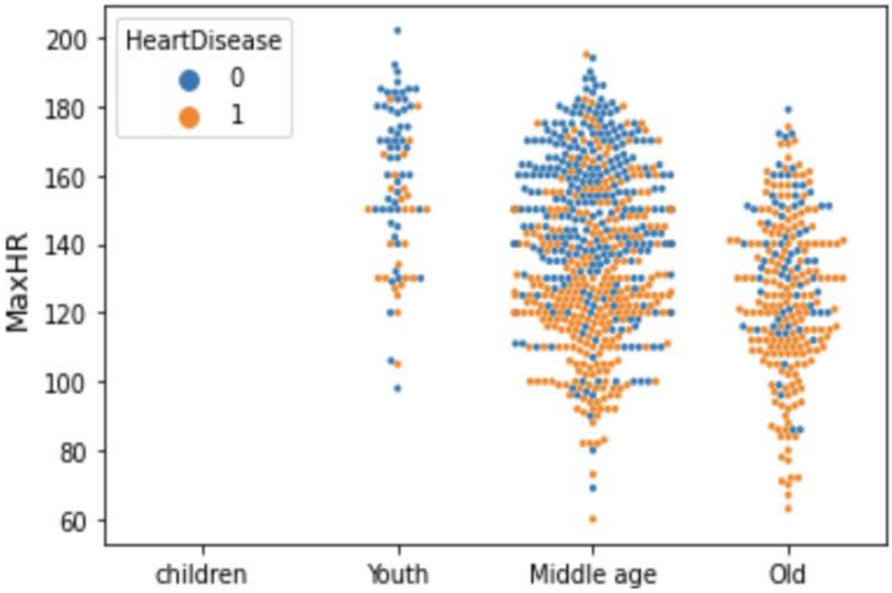

Figure 5. Tree diagram of heart disease and age.

And the analysis of the relationship between resting blood pressure and the prevalence of heart disease in Figure 6 reveals that the resting blood pressure of patients with heart failure is slightly higher than that of normal individuals. This suggests that resting blood pressure may be a risk factor for heart disease.

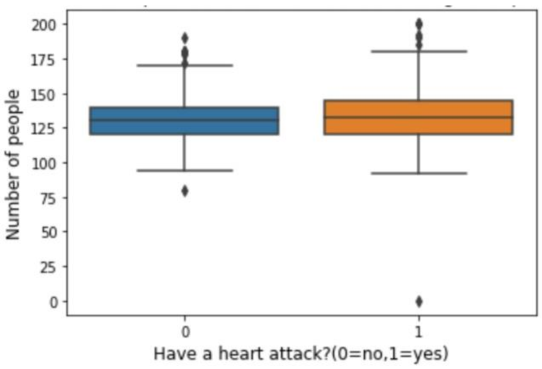

Figure 6. Relationship between heart disease and heart attack.

3.2. *Effectiveness of hyperparameters*

The impact of different hyperparameters on the results can be analyzed based on the visualization charts presented earlier. Specifically, to investigate the effects of different class weight values, C values, and penalty values on the training and test scores of the logistic regression model. It is worth noting that the following results are obtained without using the first principal component as a feature.

Table 1 demonstrates the effectiveness of various class weight and C in linear regression model.

For class weight, it can be observed that the model performs better without any class weight (train score: 0.864, test score: 0.887) than with a balanced class weight (train score: 0.856, test score: 0.883). This suggests that the imbalance in the dataset is not severe enough to require balancing the class weights.

The results indicate that the model performs best with a C value of 0.1, as evidenced by a higher test score (0.887) and a slightly lower train score (0.860) compared to other values of C such as 0.01, 1, 10, and 100. Therefore, it can be concluded that 0.1 is the optimal value of C for this particular logistic regression model.

Table 1. Effectiveness of different hyperparameters.

Class weight	C	penalty	Train score	Test score
None	0.01	L2	0.8634	0.9043
None	0.1	L2	0.8605	0.8870
None	1	L2	0.8648	0.8870
None	10	L2	0.8648	0.8870
None	100	L2	0.8648	0.8870
Balanced	0.01	L2	0.8561	0.8826
Balanced	0.1	L2	0.8590	0.8913
Balanced	1	L2	0.8590	0.8913
Balanced	10	L2	0.8590	0.8913
Balanced	100	L2	0.8590	0.8913

4. Discussion

In this study, machine learning technique is implemented to predict heart failure in patients. The results indicate that the logistic regression model can accurately predict heart failure with an overall accuracy of 88.7%. However, the study also has several limitations that need to be considered.

The number of samples, as one of the limitations, was relatively small, which may not adequately represent the entire population. Additionally, the dataset is collected from one hospital, and the generalization capacity to other institutions was not verified. Further investigations with broader and more inclusive participant samples are necessary to confirm and substantiate the conclusions drawn from this research.

Another limitation of the study is that the possible effects of ethnicity or race on heart failure prediction is not included. This could be an important factor to consider, as heart failure prevalence and risk factors may vary across different ethnic and racial groups. Future studies should aim to include more diverse patient populations to determine if race or ethnicity affects heart failure prediction accuracy.

Moreover, while the logistic regression model achieved good accuracy, it is possible that other machine learning algorithms could provide even better results. Future studies can explore the use of other machine learning algorithms and compare their performance with the logistic regression model.

Finally, the results of this study are limited to the specific features used in the logistic regression model. Future studies could explore additional features that could improve the accuracy of heart failure prediction.

In conclusion, while the study provides promising results for heart failure prediction using machine learning, it is important to recognize its limitations. Further studies are needed to validate the findings, address the limitations, and optimize the use of machine learning algorithms in the field of predicting heart-related diseases.

Absolutely, when analyzing the results obtained from machine learning techniques such as logistic regression, it is crucial to take into account the broader context in which the study is situated. As introduced in the paper’s introduction section, the problem of predicting heart disease is of great significance due to the high mortality rates associated with the condition. Therefore, developing accurate and reliable methods for early detection and heart disease diagnosis is essential in the field of medical research.

Regarding the methods used in this study, logistic regression is a well-established and widely used technique for binary classification problems. Based on the dataset analysis and the implementation of algorithm, this study aimed to predict heart failure using logistic regression. The results show that the logistic regression model can predict heart failure with a high accuracy rate of around 88%.

Different hyperparameters were analyzed to evaluate their effect on the logistic regression model’s performance. The results suggest that the model performs better without any class weight, with C=0.1, and with L2 penalty. These findings can guide future studies in selecting appropriate hyperparameters for logistic regression models in predicting heart failure.

5. Conclusion

In conclusion, this study demonstrates the application of logistic regression in predicting, a dataset of medical records was leveraged for constructing algorithms to automatically evaluate the risk of a patient suffering from heart disease. The logistic regression model achieved a high level of accuracy in identifying the risk factors and in predicting the probability of heart disease. The model holds promise for being employed as a screening method for identifying heart disease in its early stages, thereby enhancing overall public health outcomes. Future research should focus on improving the dataset’s representativeness and expanding the sample size to enhance the model’s generalizability. By doing so, future work can develop more accurate and reliable methods for detecting and diagnosing heart disease, potentially saving countless lives.

References

[1] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., et, al. (2022). Heart disease and stroke statistics—2022 update: a report from the American Heart Association. Circulation, 145(8), 153-639.

[2] Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., et, al. (2019). Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation, 139(10), 56-528.

[3] Fuchs, F. D., & Whelton, P. K. (2020). High blood pressure and cardiovascular disease. Hypertension, 75(2), 285-292.

[4] Nazarzadeh, M., Pinho-Gomes, A. C., Byrne, K. S., Canoy, D., Raimondi, F., et, al. (2019). Systolic blood pressure and risk of valvular heart disease: a Mendelian randomization study. JAMA cardiology, 4(8), 788-795.

[5] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[6] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. IEEE Access, 8, 107562-107582.

[7] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[8] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 9, 381-386.

[9] Amarnath, B., Balamurugan, S., & Alias, A. (2016). Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset. Journal of Engineering Science and Technology, 11(11), 1639-1646.

[10] LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399.