Influencing Factors and Prediction of Heart Disease

Changhai Xia *

College of Science, Shanghai University, Shanghai, 200444, China

* Corresponding Author Email: 1814010114@stu.hrbust.edu.cn

Abstract. Heart disease (HD) is a key issue of concern in the global safety and health field nowadays. Some researchers listed the predisposing factors and established some models to predict the incidence rate. However, there is still a gap in the research on the influence of related factors and the comparison of different models. Therefore, the research topic of this article is the importance of HD factors and the comparison of model predictions. Following are this article's research methods: Firstly, this project preprocesses and visualizes the data based on the Heart Failure Prediction Data Set. Then this article establishes three mainstream research models, including Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) model. Finally, this article introduces the Receiver Operating Characteristic (ROC) curve for better comparison of models and illustrates the importance of variables on HD through the ranking graph. Research has found that the RF model has the highest prediction accuracy, reaching 86.34%, and its Area Under Curve (AUC) is also the largest. Therefore, through comparison, the predicted accuracy of the RF model is the best. Meanwhile, in the ranking image of variables, it can be concluded that the Slope of peak exercise ST segment (ST_Slope) has the greatest impact on HD. Due to the possibility of prediction errors and limited applicability, it is possible to consider improving the variables to enhance model accuracy and contribute to health.

Keywords: Heart Disease, Logistic Regression, Accuracy, Random Forest, Support Vector Machine.

1. Introduction

According to estimates from the WHO, heart disease (HD) causes around 12 million deaths globally each year. In the US and other developed nations, cardiovascular illnesses account for half of all deaths. It is the main cause of adult mortality as well as the primary cause of death in many underdeveloped nations. All the various heart-related illnesses are included under the term HD. In several nations, HD is the primary cause of fatalities; especially in the US, it claims a life every 34 seconds [1].

Many academics have tested the accuracy using various databases and factors. With remarkable success, researchers have been examining the use of the Decision Tree (DT) technique in the diagnosis of HD. Sitair-Taut investigated the application of DT and Naive Bayes for the detection of coronary HD using the Weka tool [2]. Tu diagnosed HD by using the Weka tool's bagging algorithm and comparing it to the DT [2]. A Random Forest (RF) algorithm efficiently diagnoses HD's decision-making process. HD is anticipated based on the likelihood of decision support. Tu concluded that the DT works effectively and that, on occasion, its accuracy is comparable to that of Bayesian classification. Vijiyaraniet completed a study titled "An Efficient Classification Tree Technique for HD Prediction" in 2013 [3]. Srinivasan explains various machine learning (ML) algorithms. To identify the best feature set inside the data set, he described several feature selection procedures, such as Rough Set Analysis. Additionally, he used other machine learning methods, such as Naive Bayes and Support Vector Machine (SVM) [3].

Medical diagnosis is thought to be an important but challenging undertaking that requires precision and effectiveness. This system would greatly benefit from automation. However, not all subspecialties are covered by doctors, and certain locations lack enough resource personnel. Therefore, by combining them all, an automatic medical diagnosing system would be quite helpful. Clinical testing and cost reduction can be achieved with the help of suitable web-based information and decision-support tools. A comparative analysis of different approaches is necessary for the accurate and successful implementation of automated systems [4].

Based on the Heart Failure Prediction Data Set, three models including LR, RF, and SVM are established respectively to predict the incidence rate. This study aims to provide and compare effective methods for early prevention of the disease and the development of medical treatment plans.

2. Data Processing and Visualization

By referencing the data set, the goal of this project is to identify physiological factors that exacerbate HD and how they relate to HD. At the same time, different models can be established to predict given data, and the model with the best prediction effect can be determined through comparison. Considering that the data set may have some incompleteness or variables need to be transformed, before creating the model, the data must be preprocessed.

The data set should be described initially. Heart Failure Prediction Data Set is the name of the Kaggle data collection. The data set is produced by merging many datasets that were previously available separately but had never been integrated. There are 12 columns and 918 observations in the set. Additionally, it is the largest heart data collection currently available for research because it combines five cardiac data sets—including the well-known Cleveland and others—over eleven common properties [5].

Before the visualization, it is necessary to check whether there are missing values. After the examination, there is no missing value. Besides, there are six categorical variables, including Sex and so on, which cannot be processed as character strings or continuous numeric data by R language. Therefore, they need to be factorized.

The next step is to check outliers on variables. Focusing on relationships among the variables, all of them appear to be normal except Resting blood pressure (RestingBP) and Cholesterol level having outliers at 0, and Measurement of ST in depression (Oldpeak) having an abnormal number of data at 0 as well.

Then, the outliers should be eliminated. Firstly, for Oldpeak, the variable is compared to the Slope of the peak exercise ST segment (ST_Slope) after the observation. The ST segment is a description in the electrocardiogram report. ST_Slope variable has three conditions, including Up for upward-sloping, Flat for flat, and down for downward-sloping. The number of zeros would make sense as the number of Flat values in ST_Slope is the highest. Secondly, for RestingBP, observation with RestingBP=0 should be filtered out. Then, the RestingBP is compared to other variables except for Cholesterol, and it is seen that the only variable that has a noticeable difference is Exercise-induced angina (ExerciseAngina), so the median value of ExerciseAngina=No is used to impute the value [6]. Thirdly, for the Cholesterol, after the comparison, the most significant predictor would be the Sex variable. Therefore, the median cholesterol level based on Sex is used.

After the processing, the comprehensive chart in Figure 1 is drawn to depict the relationship between HD and other variables. Figure 1, green represents no HD, and red represents HD.

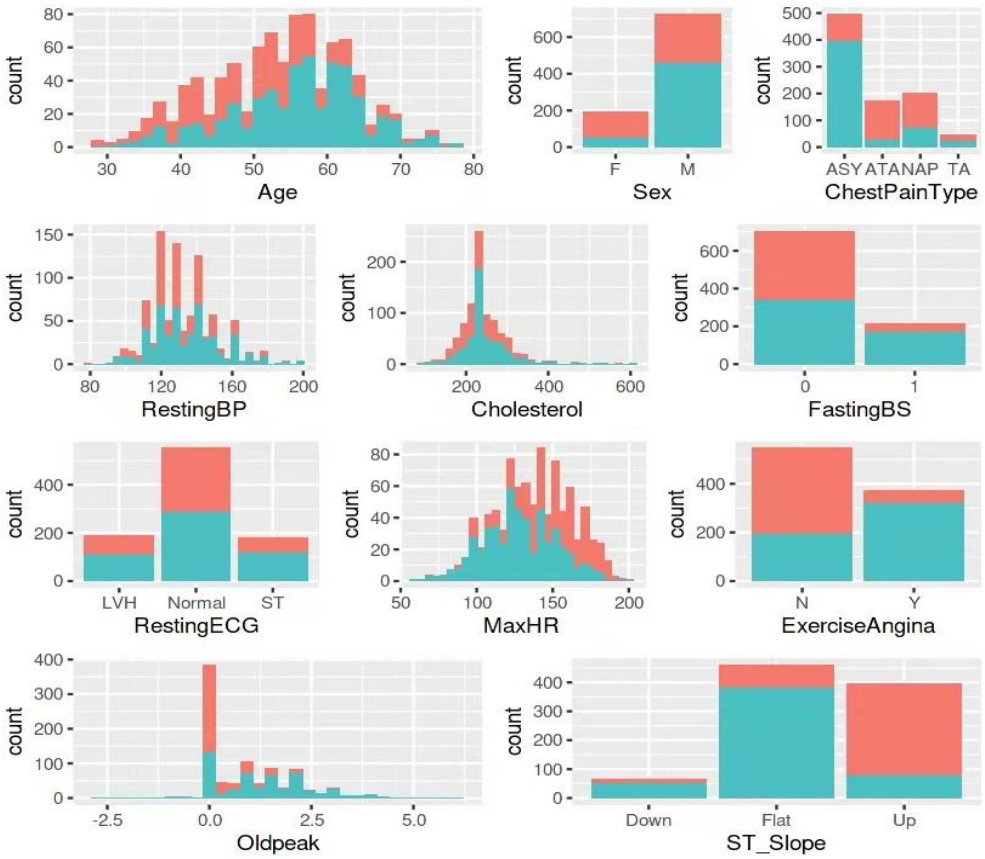

Figure 1. Visualization of processed data variables [2].

Finally, comparing the HD results with all numerical variables, the correlations can be visualized in Figure 2.

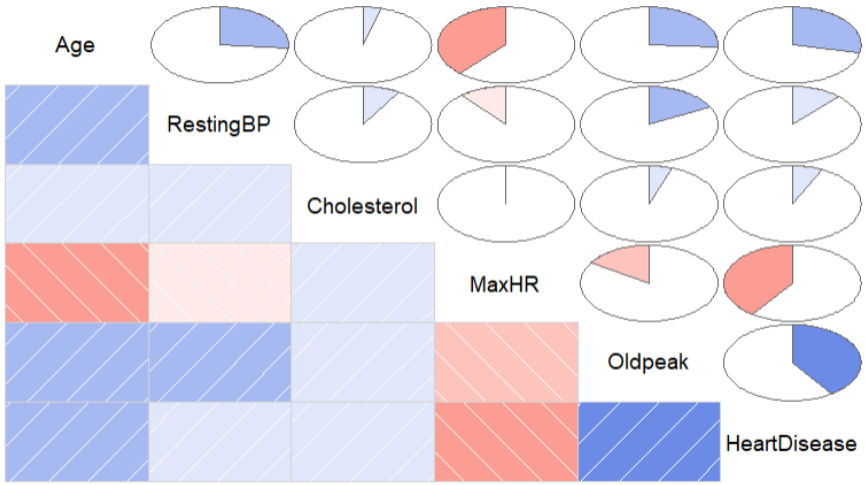

Figure 2. Correlation between numeric variables (Picture credit: Original).

From Figure 2, it appears that Oldpeak and MaxHR have significant correlations in determining whether a person has a high risk for HD.

3. Building Several Models

The models in this program include the LR Model, RF Model, and SVM Model. Before model building, it is necessary to divide the data set into training and testing sets. 80% comes from the training set, while 20% comes from the testing set.

3.1. LR Model

The extended linear regression analysis model, or LR, is a component of supervised learning in ML. Although the derivation process and calculation approach resemble the regression process, they are primarily employed to answer binary classification questions, albeit they can also address multi-classification questions.

Suppose there is a binary classification problem that requires predicting which of two categories a sample belongs to. The LR model initially determines the output probability of a given input variable using a parameterized function known as the sigmoid function (SF). It maps the linear combination of input variables to values between 0 and 1, signifying the likelihood that the anticipated sample falls into the positive scenario. The next step is model training, which involves estimating the weights of the model by maximizing the likelihood function (LHF). The LF is a function of model parameters that represent the probability of a sample given a model. A gradient descent algorithm (GDA) can be used to update the model's weights to maximize the LHF. GDA minimizes the loss function (LF) through repeated iterations until the optimal solution is found, and the LF is usually measured by the logarithmic LF to assess the model's performance. Following the model's training, it can be used to predict the class labels of new samples by substituting the feature vectors of the new samples into the SF and calculating the output probability. It is projected to be a positive case if the likelihood is greater than 0.5, and a negative case otherwise [7].

3.2. RF Model

RF is a relatively new ML model (non-linear tree-based model) ensemble learning method. Its findings are resilient to missing and unbalanced data, and it is not affected by multicollinearity. It is regarded as one of the greatest algorithms available today and has the ability to accurately forecast the impact of hundreds of explanatory factors [8].

The following are the basic steps of RF. The first step is to randomly select data. To create a sub-dataset (SDS), a replacement sampling is first collected from the original data set. The SDS contains the same quantity of data as the original data set, and it is possible to replicate elements from other SDSs. Building a sub-decision tree (SDT) with an SDS and inserting this data into it is the second stage; each SDT produces a result. Voting on the SDT's judgment results can yield the RF's output if further data needs to be categorized using an RF. The random selection of potential features is the third phase. Instead of using every candidate feature at every node, the RF SDT chooses a certain number of features at random from among all the candidate features and then chooses the best feature from the features that were chosen at random. This can increase system variety, differentiate the decision trees in the RF, and improve classification performance. The importance scoring is the fourth step. Each decision tree uses outsourcing data (OD) for prediction and records the prediction error of the OD, then creates a new OD by randomly transforming each predictor variable and utilizing the OD for validation. Finally, the mean difference between the converted prediction error and the original prediction error is used to determine the significance of a given predictor variable [9].

3.3. SVM Model

Based on the Statistical Learning Theory, SVM is a novel machine learning technique. The theory uses the structural risk minimization criterion, which increases the model's capacity for generalization, reduces sample point errors while minimizing structural risk, and is not constrained by the dimensionality of the data. The model takes the classification surface at a distance larger than the distance between the two categories of samples when performing linear classification. To convert nonlinear classification into a linear classification problem in high-dimensional space, high-dimensional space transformation is utilized [10].

The SVM algorithm can be divided into the following steps. Firstly, data preprocessing entails feature scaling (standardizing the data) and splitting the data set into training and testing sets. Then, the step is to reconstruct the model, select appropriate kernel functions and penalty coefficients, and construct an SVM model. The model must then be trained using the training set, and by maximizing

the interval, the ideal hyperplane will be found. Using the trained model to generate predictions on the test set is the final stage.

4. Result

In Table 1, the testing set is used to present the accuracy of three models. It should be noted that for the SVM model, linear kernel and radial kernel are used respectively. The kernel function is the key to SVM. Different kernel functions will produce different SVM algorithms according to SVM theory. Besides, the radial kernel performs better than the linear kernel in this data set.

Table 1. Accuracy results of three models

Model	Accuracy
LR	0.8361
RM	0.8634
SVM	0.8470

Comparing the accuracy of the three models above, the RF Model has the highest accuracy. Meanwhile, the ROC curve can be drawn to present the models' properties. As shown in Figure 3, blue refers to RF, green represents LR, and red means SVM. The curve of the RF Model is near the upper right corner, and its AUC is larger than the others. Therefore, it proves that the RF Model has the best properties among the three models.

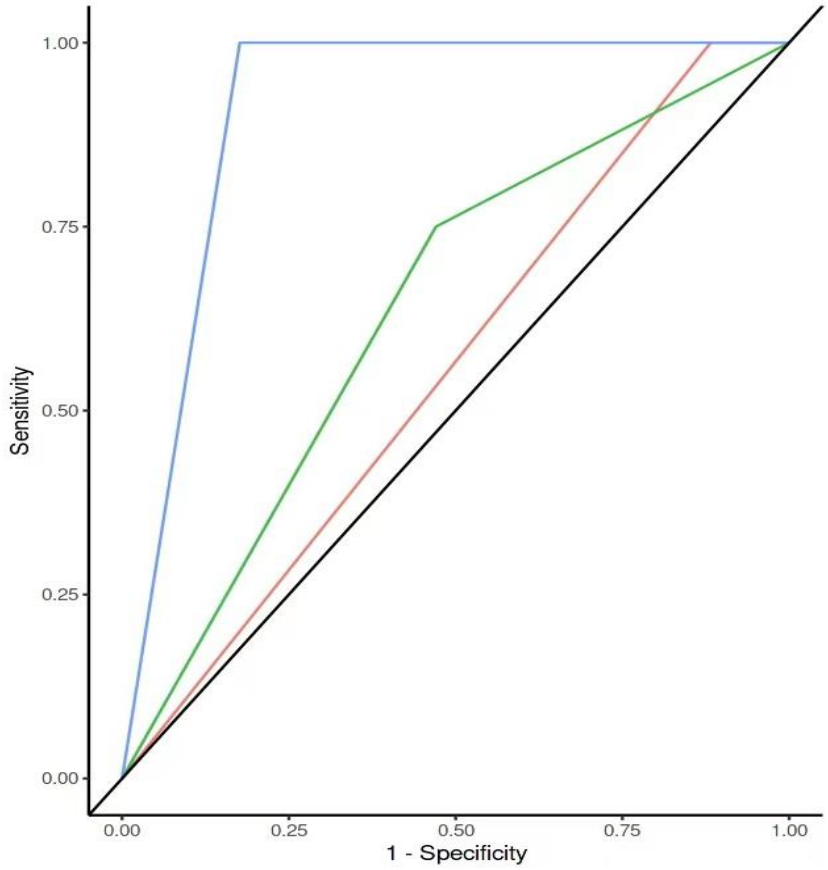

Figure 3. ROC curve of three models (Picture credit: Original).

The variables in the model can be ranked to observe which affects the model the most. As seen in Figure 4, the four most significant variables are Slope of peak exercise ST segment, Chest pain type, Oldpeak, and Maximum heart rate achieved, their amounts of importance also account for over 50% respectively.

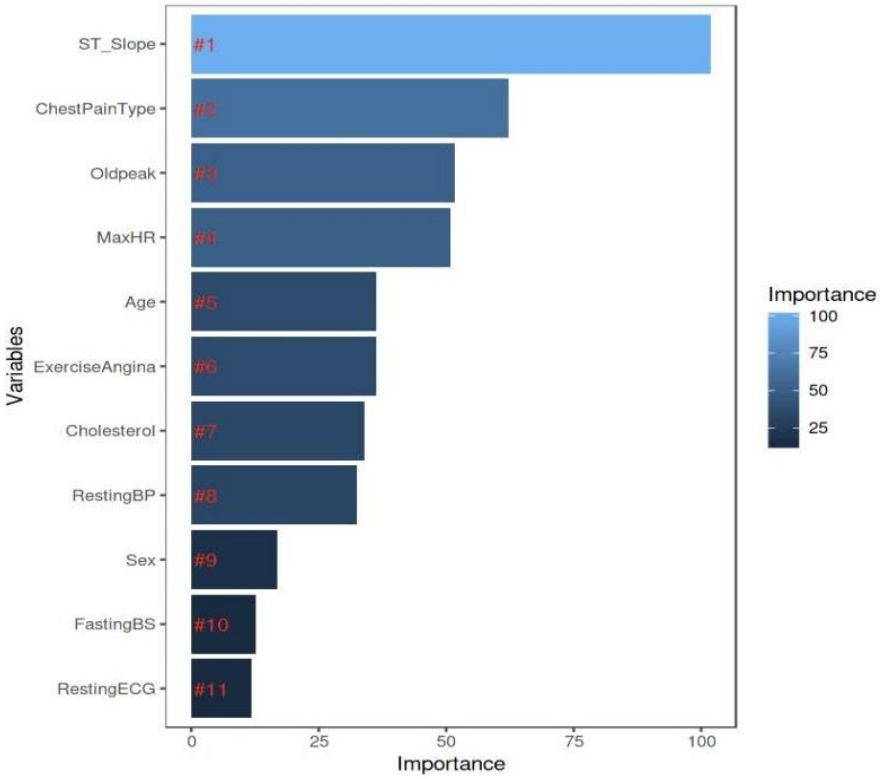

Figure 4. The ranking of HD variables [3].

5. Discussion

Considering all the factors, the RF Model is the best among them, because it can handle high-dimensional data well and is insensitive to missing values. However, the accuracy of the model can be improved a step further. The reason for the decrease in accuracy may be that RF Models have overfitting in certain classifications, and attributes with multiple value divisions can have a significant impact on the model. One of the methods is to reconsider the variables, for some of them may not be suitable for this model, and there may be some new variables that have relationships with the model.

For the data set, it is a comprehensive set compared to before, and the number of subjects is enough for the project. However, the effects of variables in different times and spaces may be different. For example, during the pandemic, the COVID-19 virus will aggravate the incidence rate of HD, which may be a major cause of the disease. Therefore, it is difficult to put into practice to predict people from different areas and episodes, and the accuracy of prediction may be affected as well.

6. Conclusion

Based on the heart failure prediction dataset, this article first preprocesses the data and completes data visualization. Then LR, RF, and SVM Models are established respectively and used to predict the heart incidence rate. Subsequently, the corresponding model accuracy was obtained, and the ROC curve and importance ranking of factors affecting HD were also displayed. Finally, the article discusses the process and results. The accuracy of the models is all above 80% and the RF Model reaches the highest. The RF Model has many advantages, including the capacity to deal with high-dimension data. However, the model is not suitable for the condition of multiple value divisions and it may cause overfitting. Meanwhile, the data set is greatly improved, but it still needs to be more practical to be implemented in different regions.

This project must be continuously improved in the future from two perspectives: improving the accuracy of the model and collecting a larger range of data. Therefore, the project can provide more

accurate cases for medical treatment, and make contributions to reducing human heart incidence rate as soon as possible.

References

[1] Titti R R, Pukkella S, Radhika L S T. Augmenting HD prediction with explainable AI: A study of classification models. Computational and Mathematical Biophysics, 2024, 12 (1).

[2] Darrab S, Broneske D, Saake G. Exploring the predictive factors of HD using rare association rule mining. Scientific Reports, 2024, 14 (1): 18178 - 18178.

[3] Mienye D I, Jere N. Optimized ensemble learning approach with explainable AI for improved HD prediction. Information, 2024, 15 (7): 394 - 394.

[4] Islam A M, Majumder H Z M, Miah S M, et al. Precision healthcare: A deep dive into ML algorithms and feature selection strategies for accurate HD prediction. Computers in biology and medicine, 2024, 176108432 - 108432.

[5] A. H A A, Prabu P, Chandra R P, et al. Comprehensive evaluation and performance analysis of ML in HD prediction. Scientific Reports, 2024, 14 (1): 7819 - 7819.

[6] Omar D E, Mat H, Karim A Z A, et al. Comparative analysis of LR, Gradient Boosted Trees, SVM, and RF algorithms for prediction of acute kidney injury requiring dialysis after cardiac surgery. International journal of nephrology and renovascular disease, 2024, 17197 - 204.

[7] Yang Yanping, Li Rong Factors affecting HD based on Decision Tree LR model. Industrial control computer, 2024, 37 (08): 114 - 116.

[8] Sridharan K. An automated HD prediction approach using linearly support vector regression and stacked linear swarm optimization. Journal of Intelligent & Fuzzy Systems, 2023, 44 (2): 3189 - 3202.

[9] Mert O, Serhat P. A classification and regression tree algorithm for HD modeling and prediction. Healthcare Analytics, 2023, 3.

[10] Hendriks M P, Bosch D V E A, Geenen W L, et al. Blood biomarkers predict 10-year clinical outcomes in adult patients with congenital HD. JACC. Advances, 2024, 3 (9): 101130.