DOI: 10.54254/2755-2721/120/2025.18749

Heart Disease Prediction Based on Machine Learning

Songze Li[1,a,*]

[1]School of Computer Science and Engineering, Xi'an Technological University, Xi'an, China
a. lisongze@st.xatu.edu.cn
**corresponding author*

Abstract: In recent years, the Heart disease has become an important cause of death in the world, early detection and treatment of symptoms is a very effective method, while in recent years machine learning technology is also gotten closer to daily life, the combination of machine learning and heart disease prediction, the use of machine learning algorithms for the basic prediction of the patient's disease; through the use of Naive Bayes(NB), Support Vector Machine(SVM), Decision Tree(DT), Random Forest(RF) and K Nearest Neighbors(KNN) by using these algorithms to provide some suggestions to the patients and provide personalised treatment suggestions, so that the patients can get the results in time and have prevention of heart disease; using accuracy to evaluate the results and evaluate each model. In this experiment, the algorithms are implemented and the accuracy of the results of each algorithm is analysed by comparing them horizontally to obtain the optimal algorithm for prediction.

Keywords: Machine Learning, Heart Disease Prediction, Naive Bayes, Support Vector Machine, Decision Tree.

1. Introduction

Heart disease is a non-communicable disease of great concern globally, according to the World Health Organisation, About 17.9 million people die of cardiovascular disease each year, accounting for about 32% of all global deaths [1]. In 2021 cardiovascular disease caused 20.5 million deaths, heart disease also caused an extremely high mortality rate, in China, the prevalence of Cardiovascular disease (CVD) is still continuing to rise, the projected number of people suffering from CVD months 330 million, in the composition of the proportion of deaths of urban residents in China, CVD is still in the first place [2]. Heart disease should be based on the concept of early detection and early prevention, can use the combination of data from previous patients, to build some prediction models to classify the patient's condition prediction, can be in a timely manner to their heart disease prevention in advance.

Ma Liyuan et al. analysed the cardiovascular multifaceted content [3]. Tie Zheng Sun and Zehao Yu studied with Cleveland database and showed that KNN algorithm performs best in this dataset and pointed out that resting blood pressure, etc. is an important indicator [4]. Syeda Urwa Warsi et al. proposed a hybrid GA - SVM method with an accuracy of 98.0% [5]. The plain Bayesian method of Anjan Nikhil Repaka et al. performed well both in terms of accuracy and performance [6]. Tejaswi. Borra et al. reviewed heart disease prediction system and discussed about it [7]. Kanksha Kaur et al. evaluated multiple algorithms using FHS dataset and SVM accuracy was up to 86.53% [8]. Xin Wang studied with UCI database and showed that Random Forest algorithm predicts best in its dataset [9]. U. Ravi Teja et al. system based on Random Forest algorithm has 95% classification accuracy on standard symptom dataset [10]. Shanshan Zhou and Yun Chen elaborated on the progress of AI application in cardiac disease diagnosis and treatment and the challenges and future directions [11], such as data integration and sharing and other trends will drive change, but also need to pay attention to the limitations and ethical and legal issues. In this paper, the accuracy of different predictions of heart disease is compared using cross-sectional algorithm accuracy comparisons, combined with datasets on algorithm implementation.

2. Research methodology

2.1. Introduction to the dataset

This study uses the Cleveland dataset, the Cleveland Heart Disease dataset is data collected by cardiologists at the Cleveland Heart Clinic in the United States and stored in the UCI Machine Learning Repository, and contains 303 samples.

Table 1: Dataset and Attributes.

Attributes	Style	Data Range
age	Continuous	29-77 years old
sex	Discrete	0 for female, 1 for male
cp	Discrete	0 indicates typical angina, 1 indicates atypical angina, 2 indicates non-anginal, and 3 indicates asymptomatic
trestbps	Continuous	94-200Hg
chol	Continuous	126-564mg/dL
fbs	Discrete	0 means ⩽120mg/dL,1 means ⩾120mg/dL
restecg	Discrete	0 indicates normal, 1 indicates ST-T wave abnormality, 2 indicates definite left ventricular hypertrophy
thalach	Continuous	71-202b/min
exang	Discrete	0 means no, 1 means yes
oldpeak	Continuous	0-6.2mV
slope	Discrete	0 means up, 1 means flat, 2 means down
ca	Continuous	0-4
thal	Discrete	0 means missing, 1 means normal, 2 means fixed defects, 3 means reversible defects
target	Discrete	0 means no, 1 means yes

The 14 attributes of the dataset are shown in Table 1, of which only age, resting blood pressure, body cholesterol, maximal heart rate, exercise relative to for rest-induced ST-segment depression, and number of major blood vessels are continuous variables, and the rest are discrete variables, are shown below in the graph of correlation analysis between factors in the analysed dataset.

Figure 1 is the correlation matrix, showing the correlation between these variables.

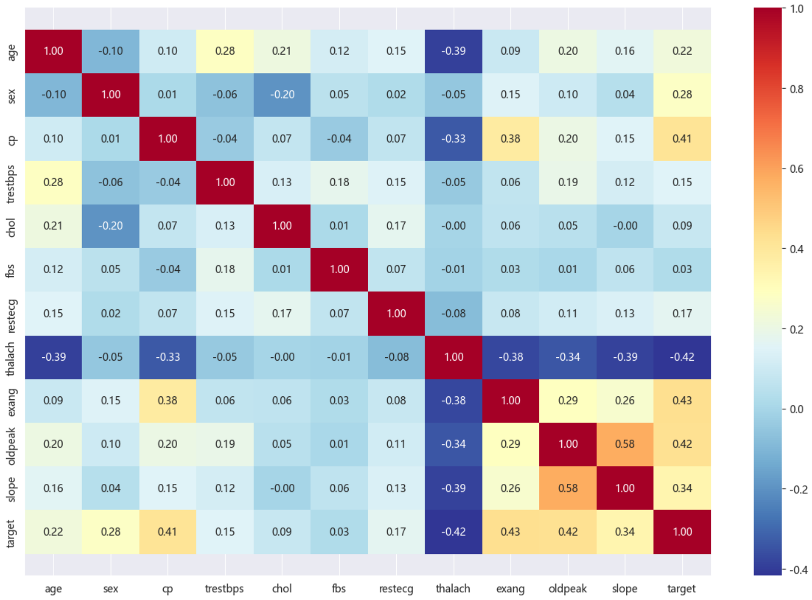

Figure 1: Correlation matrix.

2.2. Methodology

The algorithms are implemented to combine the data set, predict and analyse the attributes, compare the results with the results of the illnesses given in the data set, compare the results, and adjust the number of rounds of model training and the proportion of the data set in order to get the optimal results.

Naive Bayes (NB): NB is an algorithm for supervised learning, where the outcome is known (a priori probability), and the outcome is multiplied by the phenomenon that occurs conditional on that outcome (conditional probability) to the joint probability of the outcome and the phenomenon occurring at the same time. Dividing by the probability of the phenomenon occurring alone gives the probability of the outcome occurring conditional on the phenomenon occurring (posterior probability)

Support Vector Machine(SVM): A binary classification model is designed to discover a hyperplane for segmenting the samples, and the principle of this segmentation is to maximize the interval. The goal of SVM is to find this hyperplane.

Decision Tree(DT): Split the dataset into smaller subsets according to the different values of the data features, and select the optimal features at each split point. The computational complexity isn't overly high. The output result is straightforward to comprehend. It isn't highly sensitive to the absence of intermediate values and is capable of handling irrelevant feature data.

Random Forest(RF): RT improves the accuracy and robustness of the model by constructing multiple decision trees and aggregating their predictions. For unbalanced datasets, it balances the error; if a large portion of features are missing, accuracy can still be maintained.

K Nearest Neighbors(KNN): KNN determines which category the new sample belongs to based on what the K nearest sample points are to the target sample. Easy to use, fast training time, good prediction, insensitive to outliers.

3. Result

3.1. Evaluation indicators

Prediction calculations are performed by the algorithm and the Cleveland dataset, while the Accuracy, Precision, and Recall of the algorithm's prediction results, with accuracy being the main focus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The number of true positive cases (TP), false positive cases (FP), true negative cases (TN), and false negative cases (FN).TN denotes the number of non-diseased samples that the process correctly predicted to be non-diseased; FP denotes the number of non-diseased samples that the process incorrectly predicted to be diseased; FN denotes the number of diseased samples that the process incorrectly predicted to be non-diseased; and TP denotes the number of diseased samples that the process correctly predicted to be number of diseased samples.

3.2. Analysis of experimental results

Naive Bayes algorithm showed a relatively high accuracy rate of 87% as shown in Figure 2, demonstrating its effectiveness in this classification task.

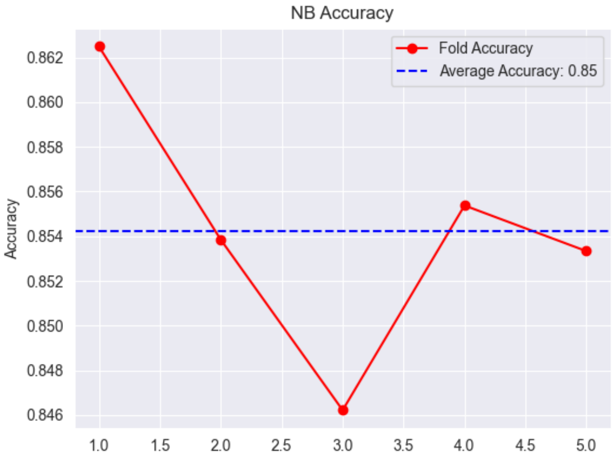

Figure 2: Naive Bayes Accuracy

Figure 3 demonstrates that the SVM algorithm has an accuracy of 81% and that the factor that the support vector machine is a binary classification algorithm affects its accuracy.

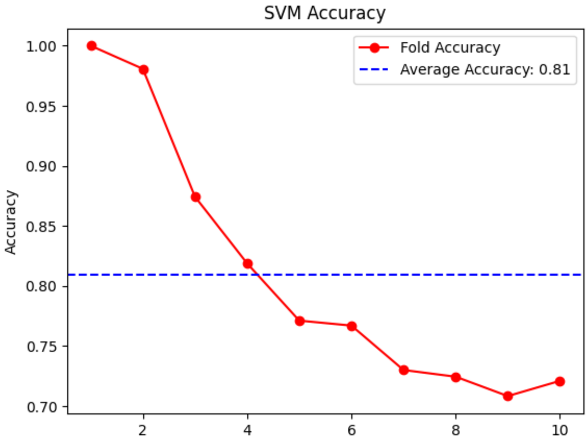

Figure 3: SVM Accuracy

Figure 4 demonstrates the accuracy of the decision tree algorithm, which is relatively low at 74%, and its prediction process is more homogeneous and unable to avoid some computational chance.

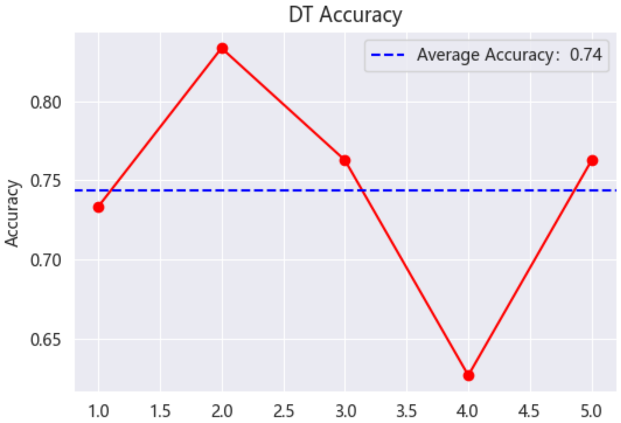

Figure 4: DT Accuracy

Figure 5 demonstrates the accuracy of the Random Forest algorithm, compared to the decision tree which combines multiple results and reduces the computational chance Random Forest accuracy is 83%.

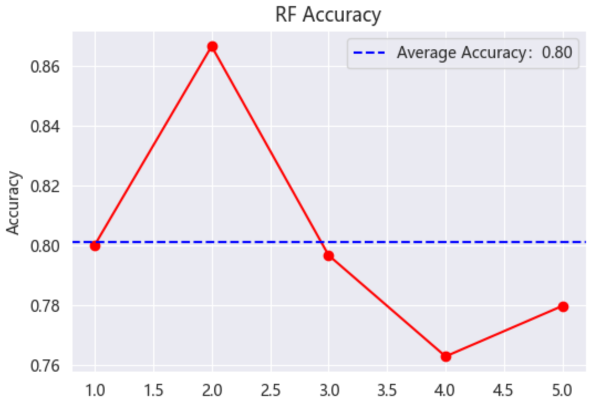

Figure 5: RF Accuracy

Figure 6 shows that the accuracy of KNN is 82%, which also verifies the accuracy of its operation, but the memory consumption is more obvious.

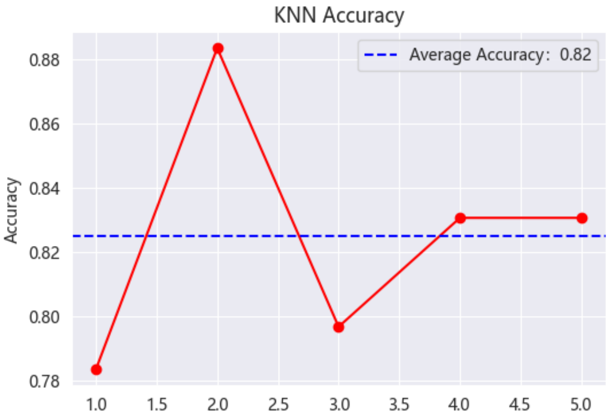

Figure 6: KNN Accuracy

4. Discussion

This paper presents an in-depth study of multiple machines learning algorithms, including Naive Bayes, Support Vector Machines, Decision Trees, Random Forests, and KNN, for heart disease prediction. The model performance was assessed by characterising and training the model on the heart disease dataset and by plotting the confusion matrix and calculating the accuracy metrics.

However, there are still shortcomings in this study, firstly, the dataset is too independent between each factor, while in reality the factors may be interdependent and influential, and then some potential relationships have to be identified when analysing the data; secondly, the samples may be predicted close to the threshold when predicting, but the algorithm judges them to be free of the disease, and there is such a monitoring prejudgement error, and lastly, for the emergence of some new premanifestation symptoms, there is not much previous empirical data for the model may not be well trained and inaccurate prediction for some symptoms.

5. Conclusion

In this study, a variety of classification algorithms were investigated and evaluated, including KNN, Naive Bayes, Random Forest, Decision Tree and SVM(Support Vector Machines). The experimental results show that different algorithms differ in classification accuracy. Among them, the plain Bayes algorithm showed a relatively high accuracy rate of 87%, demonstrating its effectiveness in this classification task. The SVM algorithm had an accuracy rate of 81%, Random Forest 83%, KNN 82%, and the Decision Tree algorithm had a relatively low accuracy rate of 74%; in the future, the algorithms will be further optimised to improve the accuracy of heart disease prediction. Meanwhile, the dataset is expanded to include more new symptom information to adapt to the changing condition. This will provide technical support for the early diagnosis and prevention of heart disease and enable people to achieve heart health.

References

[1] *Cardiovascular diseases (CVDs). Retrieved from, www.who.int/zh/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) 2024.9.2*

[2] *Li, M., Zeng, W., Jing, F., et al. (2023). Interpretation of Report on Cardiovascular Health and Diseases in China 2022. Chinese General Practice, Volume 26(Number 32).*

[3] *Tie, S., Ze, Y. (2021). Research on Classification and Prediction of Heart Cases Based on Machine Learning. Computer Knowledge and Technology, Volume 17(Number 26).*

[4] *Warsi, S. U., Mohsin, S., Asif, M., et al. (2024). A Hybrid Approach for Heart Disease Prediction using Genetic Algorithm and SVM. //2024 5th International Conference on Advancements in Computational Sciences (ICACS), IEEE.*

[5] *Repaka, A. N., Ravikanti, S. D., Franklin, R. G. (2019). Design And Implementing Heart Disease Prediction Using Naives Bayesian. //Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), IEEE.*

[6] *Borra, T., Prasad, D. G. L. V., Koyi, D. L. P. (2021). A Research Survey on State of the art Heart Disease Prediction Systems. //Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021), IEEE.*

[7] *Kaur, K., Dahiya, O., Nazir, N., et al. (2023). Heart Disease Prediction using Machine Learning. //2023 6th International Conference on Contemporary Computing and Informatics (IC3I), IEEE.*

[8] *Misra, V., Chauhan, K., Manoj, D. K., et al. (2023). Heart Disease Prediction Using Machine Learning. //2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), IEEE.*

[9] *Wang, X. (2022). Research on Heart Disease Prediction Model Based on Machine Learning. Southwest University.*

[10] *Swarupa, A., Sree, V., Sai, Y. K., et al. (2021). Disease Prediction: Smart Disease Prediction System using Random Forest Algorithm. //2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT).*

[11] *Shan, Z., Yun, C. (2023). Progress and future prospect of artificial intelligence in the field of diagnosis, treatment and research of heart. Chinese Heart and Rhythm Electronic Journal, Volume 11(Issue 1).*