Comparative Analysis of Predictive Models for Heart Disease: Evaluating Accuracy and Clinical Applicability

Hanmeng Niu

Human Ecology and education department, Columbus, America

niu.291@osu.edu

Abstract. Heart disease remains the leading cause of death globally, a reality driven by a complex interplay of genetic, environmental, and lifestyle factors. Recent advancements in technology and data science have dramatically enhanced our ability to analyze large datasets, facilitating the development of predictive models that can accurately estimate the risk of heart attacks. This capability is instrumental in preempting the onset of cardiovascular diseases and formulating preventive strategies. In this context, researchers have conducted extensive evaluations of different predictive models, scrutinizing their operational principles and accuracy. The study meticulously compared the performance metrics, specifically the precision and area under the curve (AUC), of four distinct models. It was found that each model presents unique advantages and limitations, which are reflected in their respective levels of accuracy and effectiveness in different scenarios. This analysis not only highlights the potential of machine learning in medical diagnostics but also underscores the need for ongoing refinement of computational approaches to enhance their applicability and reliability in clinical settings.

Keywords: Heart disease, Prediction models, Analysis and comparison.

1. Introduction

Heart disease continues to be the leading cause of death worldwide, significantly impacting public health systems and economies. The complexity of heart disease, influenced by genetic, environmental, and lifestyle factors, necessitates advanced methods for early detection and intervention. Recent advancements in data science and computational technology have revolutionized the approach to medical diagnostics, enabling the development of sophisticated predictive models [1]. These models leverage large datasets to forecast the likelihood of heart disease, offering a proactive strategy in combating this pervasive illness.

The use of predictive models in medicine is transforming the landscape of healthcare by enhancing disease prognosis and tailoring treatment responses. In the realm of cardiovascular health, predictive analytics provide critical insights that aid in the early detection and management of heart disease, potentially reducing mortality rates significantly. This research underscores the importance of precision and adaptability in these models to ensure their effectiveness across diverse populations and varying clinical scenarios.

This study addresses several pivotal questions:

1) How do different predictive models for heart disease compare in terms of accuracy and clinical applicability?

2) What are the specific advantages and limitations of each model when applied to real-world data?

3) Can these models be effectively integrated into clinical practices to improve early diagnosis and preventive care?

Utilizing a comprehensive dataset from the Kaggle platform, this study constructs and evaluates four different predictive models for heart disease. Each model's performance is assessed based on key metrics, including precision and the area under the receiver operating characteristic curve (AUC). The analysis not only explores the predictive power of these models but also examines their operational principles to identify how they can be optimized for clinical use. By detailing the capabilities and constraints of each model, the research aims to highlight the best practices for

integrating advanced analytics into healthcare strategies, ultimately enhancing the predictive accuracy and clinical relevance of these tools in the fight against heart disease.

2. Relevant Theories

In this section, researchers will introduce four prediction models used in this program. These four prediction models have different parameters and working theories.

2.1. Logistic regression model

Secondly, as a statistical tool, logistic regression can forecast the categorization of data using probability.

The input factors are mixed in a linear manner and converted into probability values using an sigmoid logistic function [2]. 0 for no heart disease and 1 for the presence of heart disease. For example, the higher the AUC value, the more accurate the prediction performance of this prediction model, while an AUC value of about 0.5 indicates that the discrimination ability of the entire model is low. Finally, the final AUC obtained on the test set was 0.5139, which shows that the logistic regression model has poor predictive ability in this data set [3].

$$p = \frac{1}{1+e-(a+bX)} \tag{1}$$

Equation 1: X is a factor, and p is the probability; a and b are the model parameters.

3. Random Forest

The random forest method performs very well in handling complex datasets, especially those with a large number of dimensions. The basic principle behind random forests is to utilize ensemble learning techniques. These trees then make judgments based on different features. Therefore, during the training phase, the model builds multiple decision trees [4]. In this project, the experimenters used the categorical and continuous variables in the data set to build a random forest model to make predictions. The random forest model in this project did not undergo extensive hyperparameter tuning because random forests are resilient in managing different types of data and missing values [5]. They then compared the model's predictions on the test set with the actual results to see how well it worked. As show in the Figure 1.

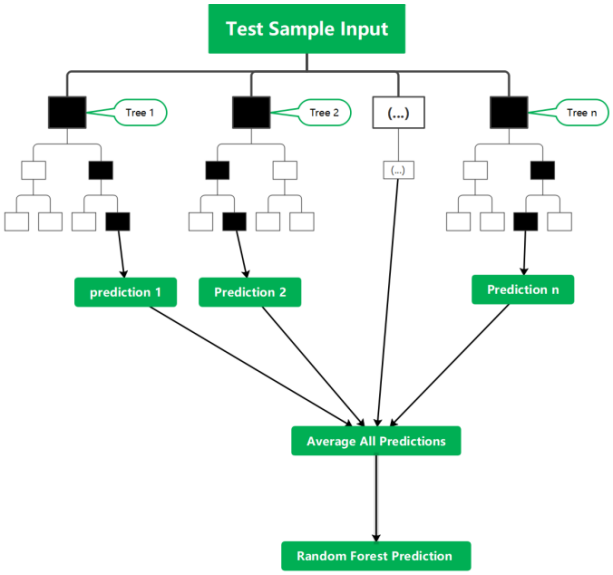


Figure 1. Random Forest Flow Chart (Photo credit: Original)

4. Decision Tree

Decision tree prediction models are often used in classification and regression applications [6, 7]. Based on the input characteristics' values, this model partitions the dataset into smaller sets. The groups are arranged in a hierarchical structure resembling a tree, with features at the root, decision rules at the branches, and outcomes at the leaf nodes [8]. Predicting the value of a target variable via data-driven learning of basic decision rules is the fundamental purpose of a decision tree [9]. The decision tree model used in this study was trained using the rpart method, which stands for Recursive Partitioning and Regression Trees, iteratively divides the data into smaller groups to maximize the homogeneity of the desired outcome within each subsequent subgroup. The splitting criteria is chosen based on which feature has the most information gain, which is assessed by the decrease in entropy before and after the split. The algorithm will continue this operation until it satisfies the predetermined ending criteria. As show in the Figure 2.

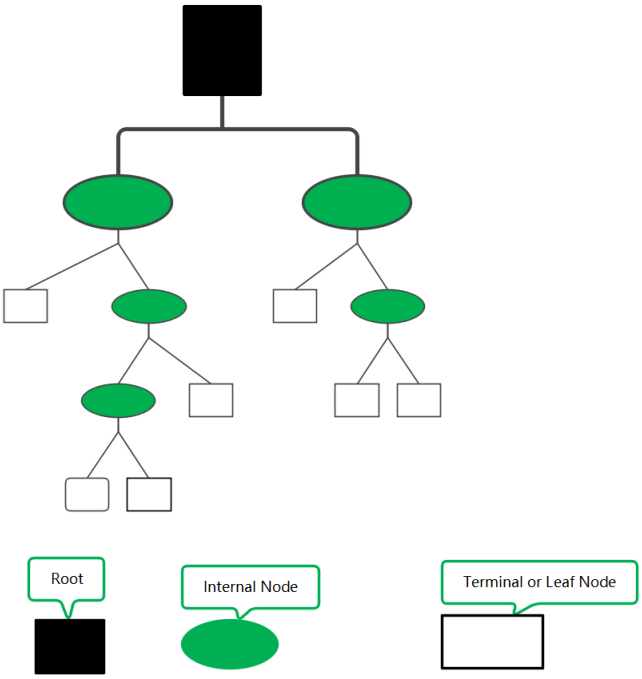


Figure 2. Decision Tree Flow Chart (Photo credit: Original)

5. K nearest neighbor

For this prediction model, first, the experimenter needs a sample data set with known labels, called the training sample set. When a new unlabeled data is provided, the algorithm examines the data by comparing unlabeled data to the data in the training sample set. Then, based on the category information of these k nearest neighbors, the category of the new data is determined by majority voting (The category with the most appearances of these k data will be classified as the input data category with feature value) [10].

The researchers in this study classified the patients into several risk categories according to their features, which were assessed using 24 criteria. A lower value of "k" might increase the model's susceptible. A higher value of "k" results in more smoothing of predictions, whereas a lower value of "k" leads to more noise. This might possibly improve accuracy by taking into account more neighbouring data points in the decision-making process.

Upon completion of the training model, the experimenters assessed the model's performance primarily based on its prediction accuracy. Nevertheless, it was ultimately demonstrated that the precision of KNN just attained a value of 0.588. Thus, in this dataset, the performance of KNN is not exceptional. As show in the Figure 3.

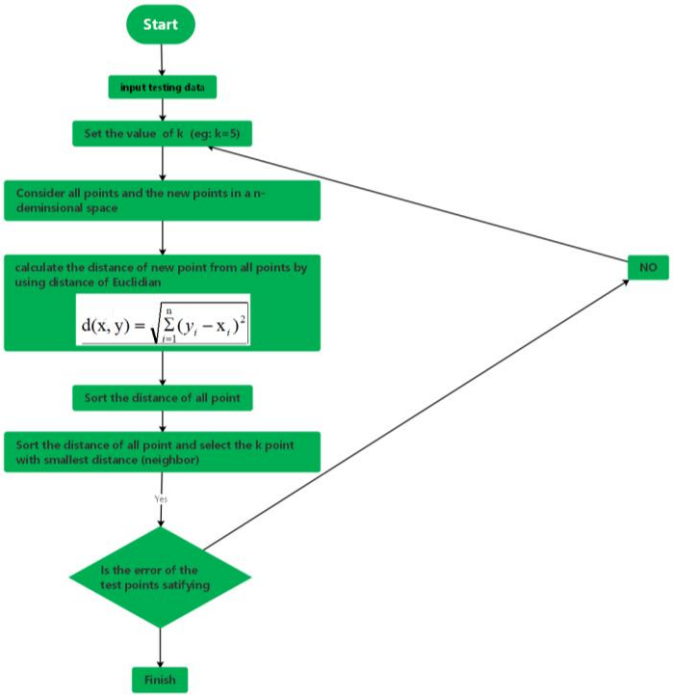


Figure 3. K nearest neighbor Flow Chart (Photo credit: Original)

6. Comparison

Then, based on accuracy and AUC, researchers can analyze the advantages and disadvantages of each model. Logistic regression has an AUC of 0.5019 and an accuracy of 0.6391. This model is restricted in its ability to address nonlinear relationships due to its assumption of a linear relationship between features and results. This is why its accuracy is medium among the four models. Random forests perform well in classification tasks in diverse data sets because they can effectively prevent over fitting (as long as a sufficient number of trees are used).

Reason being, 'k' value and distance metric selection has a significant impact on KNN's accuracy. Although this model is simple and easy to use, it may have poor computational efficiency and performance on large datasets. On the other hand, the decision tree model has a test set accuracy of 0.6476, which is relatively good, indicating that this model has better prediction results in this dataset. However, decision trees are prone to over fitting, especially when not pruned, which is particularly evident in more complex medical datasets. For example, in the application of heart disease prediction, if proper pruning is not performed, the model may capture noise in the data, thereby affecting its generalization ability. Despite this, the interpretability of decision trees in medical decision making is still a significant advantage because it can show the contribution of each decision node to the final prediction through clear rules.

When comparing the parameters of these models longitudinally, several differences can be found. The logistic regression model is simple and straightforward, with no complex hyperparameters, and relies primarily on the assumption that the underlying relationship is linear, but this feature can also limit its flexibility and performance in more complex datasets.

These parameters allow random forests to effectively manage high-dimensional data and reduce the possibility of overfitting, but at the cost of increased computational complexity. Larger ‘k’ values may provide more robust predictions, but may also blur the differences between categories, resulting in lower sensitivity. Although deeper trees can capture more details, they are also more likely to overfit in datasets containing noise or outliers.

7. Challenge

Although the experimenters used the correct logic code in building these four models, the final accuracy is still relatively low. The reason may be that in the decision tree model, due to insufficient attention to the parameters, the model did not find the best parameter value, thus reducing the accuracy. Therefore, to make the model more precise, the experimenters can use methods such as neural networks and interdisciplinary cooperation to solve the underfitting problem. Within the random forest model, researchers have the ability to modify parameters such as the quantity of trees, the maximum depth, and the minimum number of sample splits in order to enhance the accuracy. For KNN, the experimenters can try different 'k' values to improve the accuracy. But in general, the sample data of about 9,000 patients is still too small, so if the experimenters can recruit more patients from different places and different races, the established model will have a higher accuracy.

Table 1. Specific accuracy and AUC value

	Logical Regression	Random Forest	K Nearest Neighbour	Decision Tree Classifier
train_y with validation set	AUC: 0.5019 Accuracy: 0.6391	Accuracy: 0.6345	AUC: 0.4949 Accuracy: 0.5711	AUC: 0.4890 Accuracy: 0.6202
train_y without validation set	AUC: 0.5056 Accuracy: 0.6391		AUC: 0.4949 Accuracy: 0.5711	AUC: 0.4890 Accuracy: 0.6202
test_y with validation set	AUC: 0.5103		Accuracy: 0.5888	Accuracy: 0.6476
test_y without validation set	AUC: 0.5103		Accuracy: 0.5888	Accuracy: 0.6476
val_y	AUC: 0.5139		Accuracy: 0.6052	Accuracy: 0.6452

8. Conclusion

The final accuracies of the four evaluated models were as follows: Logistic Regression at 0.5103, Random Forest at 0.6345, K-Nearest Neighbors at 0.5888, and Decision Tree at 0.6476. Each model exhibits distinct strengths, but their performance and suitability vary considerably based on dataset characteristics and underlying predictive principles. For instance, the Random Forest model demonstrates superior accuracy with high-dimensional data, leveraging its ensemble approach to mitigate overfitting and enhance generalization. Conversely, Logistic Regression, despite offering clear interpretability, is constrained by its foundational linear assumptions. The Decision Tree model, although susceptible to overfitting, offers a straightforward, intuitive structure and decision-making process, making it both easy to understand and explain. This simplicity, coupled with robust performance on the dataset, marks it as the most effective model in this study. Detailed performance metrics, including accuracy and area under the curve, are illustrated in Table 1. This analysis underscores the critical importance of careful model selection and optimization in the predictive modeling of heart disease. It highlights the necessity to tailor model choice to the specific features and challenges of the dataset at hand, ensuring optimal outcomes in future healthcare applications and enhancing the precision of preventive strategies.

References

[1] J. S. Aluru, et al., "Valvular Heart Disease Epidemiology," Medical sciences (Basel, Switzerland), vol. 10, no. 2, p. 32, Jun. 2022, doi: 10.3390/medsci10020032.

[2] X. Zhu, H. Xu, Z. Zhao, X. Wang, X. Wei, Y. Zhang, and J. Zuo, "An environmental intrusion detection technology based on WiFi," Wirel. Pers. Commun., vol. 119, no. 2, pp. 1425-1436, 2021.

[3] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," International Journal of Computer Science Issues, vol. 9, no. 5, pp. 272-278, 2012.

[4] P. Ghosh, et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," IEEE Access, vol. 9, pp. 19304- 19326, 2021.

[5] X. Zhu, C. Guo, H. Feng, Y. Huang, Y. Feng, X. Wang, and R. Wang, "A Review of Key Technologies for Emotion Analysis Using Multimodal Information," Cogn. Comput., vol. 1, no. 1, pp. 1-27, 2024.

[6] A. N. V. K. Swarupa, V. H. Sree, S. Nookambika, Y. K. S. Kishore, and U. R. Teja, "Disease Prediction: Smart Disease Prediction System using Random Forest Algorithm," in 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, 2021, pp. 48-51.

[7] S. Yousefi, "Comparison of the performance of machine learning algorithms in predicting heart disease," Frontiers in Health Informatics, vol. 10, no. 1, p. 99, 2021.

[8] X. Zhu, Y. Huang, X. Wang, and R. Wang, "Emotion recognition based on brain-like multimodal hierarchical perception," Multimed. Tools Appl., vol. 83, no. 18, pp. 56039-56057, 2024.

[9] M. M. Ali, B. K. Paul, K. Ahmed, et al., "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Computers in Biology and Medicine, vol. 136, p. 104672, 2021.

[10] R. Katarya, and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," Health and Technology, vol. 11, no. 1, pp. 87–97, 2021.