Comparative Analysis of Machine Learning Models for Heart Disease Prediction

Weichi Gao
Institute of Science and Technology, Wenzhou Kean University, Wenzhou, Zhejiang, 325060, China

Keywords: Heart Disease Diagnosis, Machine Learning, Pattern Recognition.

Abstract: In recent years, heart disease has become one of the major public health problems worldwide. According to the World Health Organization, cardiovascular disease is one of the leading causes of death, especially among middle-aged and elderly people. Lifestyle changes, such as an unhealthy diet, lack of exercise, and high stress levels, dramatically increase the incidence of heart disease. This paper will compare three models, including decision trees, random forests, and Limit Gradient Lift (XGBoost), by analyzing heart disease data sets. Through the comparison and analysis of these three machine learning models, the final conclusion is that XGBoost model has the highest accuracy. Machine learning has significant advantages in the medical field, especially in the detection of heart disease. First, machine learning algorithms can efficiently process large amounts of data. Second, machine learning is able to identify complex patterns and small differences that are difficult to detect with traditional methods, thus improving the accuracy of diagnosis. In addition, machine learning is highly adaptive, with the ability to continuously optimize and improve models based on new data.

1 INTRODUCTION

Heart disease is a kind of disease that affects the heart function widely, including coronary heart disease, myocarditis, arrhythmia, heart failure and many other types (Ponikowski, 2014). These diseases are often caused by atherosclerosis, high blood pressure, diabetes, poor lifestyle habits (such as smoking, alcohol abuse, obesity, physical inactivity), and genetic factors. Heart disease is very harmful to human health and is one of the important causes of death and disability in the world (Groenewegen, 2020). The risks of heart disease vary, the most serious of which is sudden death, especially acute myocardial infarction, a heart attack, which often becomes life-threatening in a short period of time (Mata, 2014). In view of the high risk of heart disease, early detection and timely intervention are particularly important. Early detection can enable patients to take effective treatment before the disease progresses, significantly reducing the risk of sudden heart attack and death. Through early treatment, It can also reduce the probability of serious complications such as heart failure, thereby improving the living standards of patients and extending the life span of patients. Common methods for predicting heart disease risk include risk scoring systems, blood tests, imaging and genetic testing. Through these methods, an individual's risk of heart disease can be more accurately assessed, so that more effective prevention and treatment strategies can be developed to help patients take early measures to avoid further deterioration of the disease.

Traditional machine learning models offer a suite of advantages that are particularly beneficial in the detection of heart disease (Ahsan, 2022). These models are good at processing large-scale data, recognizing complex patterns, and can make high-precision predictions (Khan, 2019). Their ability to learn from historical data and adapt to new information is crucial in the medical field, where early and accurate diagnosis is paramount. Moreover, machine learning models are highly customizable and can be fine-tuned to improve their performance over time. They are also less prone to human error and bias, ensuring a more objective analysis of patient data. As a result, these models play a pivotal role in advancing cardiovascular disease management, contributing to better patient outcomes and a reduced burden on healthcare systems.

This essay will focus on predicting heart disease by comparing the performance of random forests, decision trees, and Extreme Gradient Boosting

Gao, W.
Comparative Analysis of Machine Learning Models for Heart Disease Prediction.
DOI: 10.5220/0013296800004558

In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 234-237
ISBN: 978-989-758-738-2

(XGBoost) algorithms. This study aims to identify the most effective models of early detection and prevention to contribute to ongoing global efforts for cardiovascular disease control and prevention. The following sections will be divided into: introduction, method, experiments and result, discussion and conclusion.

2 METHOD

2.1 Dataset

The dataset utilized in this study comprises 1,025 samples and 14 features (David, 2014). Each row corresponds to an individual, while the columns represent various health indicators that may influence the risk of heart disease. To handle categorical variables, one-hot encoding was applied. Initially, features such as "gender," "chest pain type" (cp), and "resting electrocardiographic results" (restecg) were represented by single values corresponding to different categories. With one-hot encoding, these categorical variables were transformed into multiple binary columns (True/False), each representing a specific category. This approach prevents the model from mistakenly interpreting categorical variables as ordinal, thereby enhancing the accuracy of the model's predictions.

2.2 Models

This essay will compare three models in total, including Decision Tree, Random Forest, and XGBoost.

2.2.1 Decision Tree

Decision trees represent a non-parametric supervised learning method utilized for both classification and regression problems (Song, 2015). This algorithm is structured hierarchically, with components including branches, root nodes, internal nodes, and leaf nodes. One of the key benefits of decision trees is their high level of interpretability. Because of its structure, it is easy to understand how predictions are made. It can provide a clear decision path that can be traced back to each prediction. The simplicity of decision trees comes with limitations - they are prone to overfitting, which means the model becomes too suited to the specific features of the training data. This will lead to poor performance on unknown data, such as predictions for new patients. In addition, decision trees are very sensitive. Even small changes can lead to completely different results, resulting in poor performance (De Ville, 2013).

2.2.2 Random Forest

The second model is a random forest. It is a frequently used machine learning algorithm that combines the outputs of multiple decision trees to arrive at a single result. Its practicality and flexibility have led to its adoption as it deals with classification and regression problems (Rigatti, 2017). Each tree is trained with random data and features, which allows the model to consider different combinations of risk factors. Random forests can effectively solve overfitting problems in decision trees because it can draw conclusions from multiple trees. This eliminates noise and makes it easier to generalize the model to new data. In addition, due to its multi-tree nature, it can rely on the collective decisions of other trees to process missing data. However, there are some drawbacks to this model. It is clear that random forests are computationally intensive, especially when dealing with large data sets or large numbers of trees. Moreover, it improves accuracy at the expense of interpretability. Unlike a single decision tree, where the prediction path is clear, the holistic nature of a random forest makes it harder to tell exactly how a particular decision was made.

2.2.3 XGBoost

The third model is XGBoost (Chen, 2016), a scalable and distributed gradient-boosted decision tree (GBDT) machine learning library. XGBoost is widely recognized for its efficiency and scalability, which makes it especially effective for processing large datasets.The algorithm's distributed nature allows it to perform parallel tree boosting, significantly reducing training time and enhancing its ability to tackle complex machine learning tasks such as regression, classification, and ranking. One of the key strengths of XGBoost is its meticulous approach to model training. It employs an iterative process that incrementally builds trees, with each new tree aimed at correcting the errors of its predecessors. This method enables the algorithm to refine its predictions with each iteration, leading to improved accuracy over time. The algorithm also incorporates regularization techniques to prevent overfitting, ensuring that the model maintains robustness even as it becomes more complex. XGBoost's ability to handle missing data is another notable feature. It is designed to be robust against incomplete data, allowing it to make reliable predictions even when certain information is absent. This resilience is particularly valuable in real-world scenarios where data integrity can be compromised. Despite XGBoost's high accuracy, it also faces some challenges. Like random forests, it sacrifices interpretability for better performance. In addition, it is difficult to find detailed operations for specific predictions. In summary, while XGBoost presents a powerful tool for enhancing prediction accuracy in heart disease detection, its use also necessitates careful consideration of its limitations. Future research and development should focus on enhancing the algorithm's interpretability and accessibility, ensuring that its benefits can be fully realized in practical applications.

3 EXPERIMENT AND RESULTS

The age distribution in the data set shows that 52.2% of the sample was made up of older adults (aged over 55 years), 42.1% was middle-aged (aged between 40 and 55 years), and only 5.7% was young (aged between 29 and 40 years). This distribution provides important insights into the potential correlation between age and heart disease prevalence. In addition, heat maps were used to visualize the linear relationship between the features and highlight the significant correlations. For instance, there is an inverse relationship between maximum heart rate and age, whereas a positive correlation exists between exercise-induced ST-segment depression and the number of major blood vessels, with the latter showing a correlation coefficient of 0.32.

To ensure efficient feature selection, a statistical method is used in the analysis that ranks features according to their relevance to the target variable, as shown in Figure 1. This process involves several key steps. First, the data set is processed to exclude the target variables and keep only the predictor variables. Feature selection techniques are then applied to identify the top 13 most important features. These features were then extracted and ranked according to their statistical significance. The results are visualized in a bar chart that shows the relative importance of each selected feature.

The testing of the hypothesis is done through the code development process of the system. First, a dataset of heart disease is imported, and a unique thermal encoding is used to convert categorical variables such as gender into numerical values, facilitating efficient processing of the model. The dataset is subsequently split into a training set, used for model training, and a validation set, employed to assess the model's performance. To enhance the model's prediction accuracy, a feature selection technique is applied to identify the most significant features, which are then used in the following model training process.

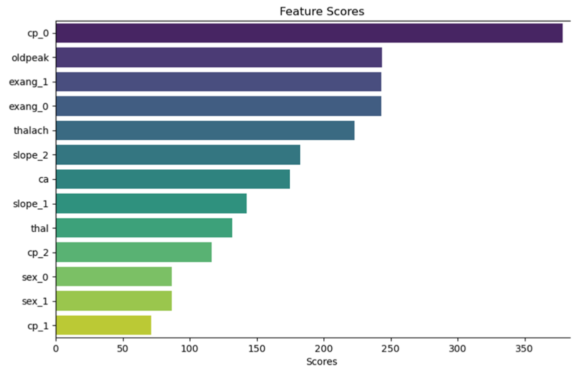

Figure 1: Importance of features (Figure Credits: Original).

Multiple decision tree models are trained using different minimum number of node split samples and maximum tree depth. The accuracy curves of the training set and the validation set were drawn to determine the optimal model parameters, and the maximum depth was determined to be 16 and the minimum sample split number to be 10. Similarly, multiple random forest models were trained and the best-performing model was selected based on its accuracy. In addition, the XGBoost model is trained on the data set and an early stop technique is introduced to prevent overfitting. The optimal number of iterations, training accuracy and test accuracy were then recorded as demonstrated in Table 1. The accuracy of decision tree, random forest and XGBoost models on validation sets is compared and analyzed by bar graph, and the most efficient models are identified.

Table 1: Accuracy comparison of different models.

Models	Accuracy
Decision tree	0.9610
Random forest	0.9659
XGboost	0.9805

4 DISCUSSIONS

It is evident that the XGBoost model achieves the highest accuracy among the models. Through analysis, it can be found that the best performance of the XGBoost model is mainly due to its advantages in ensemble learning, handling unbalanced data, efficient computing performance, and flexible tuning. These features allow XGBoost to provide greater accuracy and better generalization on complex data sets. This is why the XGboost model is more accurate than other models.

This study underscores two critical issues: the imbalance in the dataset and the narrow scope of data sources. Firstly, the age distribution's skew towards older adults not only restricts the model's generalizability to younger demographics but also potentially masks age-specific risk factors that could be crucial for comprehensive heart disease prediction. This demographic limitation could lead to underrepresentation of early-onset heart disease patterns, thereby affecting the model's predictive accuracy across all age groups.

Secondly, the reliance on a single dataset, without incorporating diverse data from multiple institutions or international sources, may introduce geographical and ethnic biases. Heart disease presents varying risk factors and manifestations across different populations due to genetic, environmental, and lifestyle differences. The lack of a multi-institutional, multinational dataset could hinder the model's ability to capture these nuances, thus limiting its global applicability and reducing its effectiveness in providing personalized risk assessments.

To address these limitations, future research should aim to develop a more balanced and diverse dataset that includes a broader age range and represents multiple populations. This strategy will improve the model's predictive accuracy and ensure it is better suited to assess heart disease risks across different demographic groups. Additionally, employing advanced feature selection techniques and dimensionality reduction methods will allow for a more holistic understanding of the complex interplay between features, leading to more accurate and nuanced predictions. Furthermore, expanding the comparative analysis to include other models like Neural Networks may reveal additional insights and potentially higher predictive accuracy. The primary objective is to improve the predictive capabilities of heart disease models, which will contribute to more effective prevention strategies and better cardiovascular health outcomes. As machine learning continues to evolve, there is great potential for developing more accurate, adaptive, and personalized tools for predicting heart disease in the future.

5 CONCLUSIONS

In summary, three models of decision tree, random forest and XGboost are compared. XGboost was found to have the highest accuracy. The decision tree is 0.9610, the random forest is 0.9659, and XGBoost is 0.9805. Therefore, it can be found that the predictions for the performance of the three models are correct. XGBoost has high accuracy. Iterative steps help it catch patterns that other models might miss, allowing it to make better predictions. In addition, it can handle lost data, which is useful in cases where records are incomplete. In the future, the model will be upgraded by employing a diverse range of models to predict heart disease cases, aiming to identify alternatives that outperform XGBoost or to further refine the existing XGBoost model for improved accuracy. Additionally, a website will be set up, with the goal of training a refined heart disease prediction model and launching a platform where users can estimate their heart disease risk and receive personalized advice on how to improve their health.

REFERENCES

Ahsan, M. M., & Siddique, Z. 2022. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289.

Chen, T., & Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785-794.

David, L. 2019. URL: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset. Last Accessed: 2024/09/09

De Ville, B. 2013. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448-455.

Groenewegen, A., Rutten, F. H., Mosterd, A., & Hoes, A. W. 2020. Epidemiology of heart failure. *European journal of heart failure*, *22*(8), 1342-1356.

Khan, Y., Qamar, U., Yousaf, N., & Khan, A. 2019. Machine learning techniques for heart disease datasets: A survey. In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 27-35.

Mata, J., Frank, R., & Gigerenzer, G. 2014. Symptom recognition of heart attack and stroke in nine European countries: a representative survey. *Health Expectations*, *17*(3), 376-387.

Ponikowski, P., Anker, S. D., AlHabib, K. F., Cowie, M. R., Force, T. L., Hu, S., ... & Filippatos, G. 2014. Heart failure: preventing disease and death worldwide. *ESC heart failure*, *1*(1), 4-25.

Rigatti, S. J. 2017. Random forest. *Journal of Insurance Medicine*, *47*(1), 31-39.

Song, Y. Y., & Ying, L. U. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.