A Comparative Study of Machine Learning Models for Cardiovascular Disease Prediction

Beining Qian
Department of Computer Science and Technology, Chongqing University, Chongqing, 401331, China

Keywords: Machine Learning, Cardiovascular Disease, Random Forest.

Abstract: The efficacy of diverse machine learning approaches in predicting cardiovascular diseases is compared in this analysis by utilizing a range of hyperparameter tuning and data preprocessing techniques to improve model performance. The methods applied include encoding categorical data, generating additional features, selecting the most relevant features, and standardizing data, along with extensive hyperparameter optimization. The models evaluated include Decision Tree, Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). The results show that compared with other models, the Random Forest Model has a unique ensemble learning method, and achieves excellent performance and robustness by virtue of this method. As demonstrated by results, the Random Forest effectively balances bias and variance and addresses the complexity of medical data. The results of the experiment proved to be the best performer in this survey was the Random Forest Model, although XGBoost also showed strong performance with its sophisticated boosting and regularization strategies. This emphasizes how crucial model tuning and selection are to improving forecast accuracy. To further improve prediction reliability and generality, future research should investigate more sophisticated models and methodologies, optimize preprocessing and tuning strategies, and incorporate larger datasets.

1 INTRODUCTION

In recent times, medical analysis has become a crucial component of modern healthcare, playing a vital role in disease diagnosis, prognosis, and treatment planning (Kononenko, 2001). With the advancement of large-scale medical data sets such as electronic health records, medical image files, and mobile apps, it has become easier to accurately analyze patient data to aid in diagnosing conditions. Medical analysis encompasses various methods aimed at understanding, predicting, and mitigating the occurrence of diseases, giving medical professionals insightful information for early intervention and individualized care (Celermajer, 2012). Machine learning is one of the most promising techniques in this field since it uses data-driven methods to identify patterns and correlations in medical information.

Machine learning approaches could intelligently learn representative feature combinations for specific missions (Jordan, 2015). It has demonstrated extraordinary promise in medical analysis for enhancing the precision of diagnoses, forecasting the course of diseases, and enhancing treatment plans. Particularly in the diagnosis of cardiovascular diseases (CVD), these models have demonstrated superior performance by identifying complex relationships between related factors such as cholesterol levels, age, blood pressure, and other clinical parameters (Mathur, 2020). Machine learning can handle non-linear correlations, manage high-dimensional data, and adjust to changes in patient populations more effectively than traditional statistical methods (Hagan, 2021).

Machine learning approaches have been leveraged in several recent research to predict disease using a variety of medical datasets. A representative work leveraged them to forecast the start of cancer, diabetes, and cardiovascular disease (Knutsson, 2000). While these studies have shown promising results, they often rely on specific feature sets, lack extensive data preprocessing, or underutilize feature engineering's potential to increase model correctness. Furthermore, it is still difficult for these models to be generalized to other populations.

This research focuses on tackling limitations in prior investigations by evaluating the performance of various machine learning techniques and conducting

Qian, B.
A Comparative Study of Machine Learning Models for Cardiovascular Disease Prediction.
DOI: 10.5220/0013332300004558

In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 377-381
ISBN: 978-989-758-738-2

an in-depth examination of their predictive capabilities for cardiovascular diseases. This study examined several data preprocessing strategies, including feature scaling, encoding, and selection, and assessed the effectiveness of models, including Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). Additionally, the author assessed the importance of various features to understand their contribution to disease prediction. Ultimately, through comparison, this work identified the most robust and interpretable prediction framework. With this research, the author strives to increase the predictability of cardiovascular illness and offer enhanced insights for predictive modeling in clinical settings.

2 METHOD

2.1 Dataset and Preprocessing

The Cardiovascular Disease Dataset, which was obtained from Kaggle, was used in this investigation. It includes 70,000 data, each representing a patient, with 11 different patient-related eigenvalues as well as a binary indicator signifying whether cardiovascular disease is present or not (Svetlana, 2018). Three categories are used to classify the features: Objective features are factual information about things like weight, height, age, and gender; examination features are results from a medical exam like cholesterol, blood pressure, and glucose levels; and subjective features are things like self-reported information about things like drinking alcohol, smoking, and physical activity. The goal variable, "cardio," is binary and indicates the presence or absence of cardiovascular illness (1) or (0). All data was collected during medical examinations, providing a snapshot of each patient's health status.

In this study, several methods are leveraged to process original data before formal experiment, in order to improve the accuracy and other evaluation Indicators. Here are those four methods: (1) Handling Missing Values: Although there are no such values in the dataset, checks were made to guarantee the accuracy of the data. (2) Categorical Encoding: One-Hot Encoding is leveraged to encode data with more than 2 categories (e.g., gender, cholesterol, and glucose in this experiment) into a numerical format for learning, i.e., 0 and 1. (3) Feature Scaling: To guarantee uniformity across various forms of data, a single scale was applied to a range of features, including age, height, weight, and blood pressure. This normalization improves the model's predictive accuracy and stabilizes its training phase. (4) Feature Selection: To enhance the model's efficiency, features with minimal variability were removed to eliminate noise and irrelevant information. Furthermore, a technique was employed to pinpoint and preserve the most revealing features based on their statistical relevance.

2.2 Models

2.2.1 Logistic Regression

It is the baseline model in this work, which is a statistical approach that is frequently leveraged for binary classification (LaValley, 2008). It introduces a linear combination of various attributes, applies logistic transformations, and constructs a Logistic Regression model to predict the probability of the target variable classifying into a particular category. In this study, it is used to predict whether an individual has cardiovascular disease (1) or not (0), using processed features. Although Logistic Regression provides a strong baseline, its performance is always not very good when the connection between the target's features variable is non-linear.

2.2.2 Decision Tree

A Decision Tree segments the dataset into subsets based on attribute values. Within this structure, each internal node corresponds to an original data feature, while each leaf node signifies an original data class label (Priyanka, 2020). In this study, it is employed to forecast the occurrence or non-occurrence of cardiovascular disease through the analysis of the refined features. The model progressively partitions the data by choosing the best split at each node, as dictated by the Gini impurity measure. Decision trees are simple to understand, but if they are not adequately managed, they may overfit the data. Adding, deleting, or optimizing features is a type of feature engineering that increases the correctness of the model.

2.2.3 Random Forest

It is an ensemble technique that relies on the use of decision trees. Its distinctive feature is the construction of numerous decision trees throughout the training phase, with the final class predictions being determined by the majority vote of the individual trees (Biau, 2016). In this study, cardiovascular illness is predicted by Random Forest using processed characteristics. This model mitigates

overfitting by averaging the predictions across multiple trees. In the Random Forest framework, each constituent tree is developed on a distinct data subset and feature set, thereby enhancing the model's precision and reliability, and thus, its overall performance.

2.2.4 XGBoost

Employing the sophisticated ensemble method known as XGBoost, a sequence of Decision Trees is built, where each subsequent tree corrects the errors of its predecessor (Chen, 2016). In this study, XGBoost uses processed features to predict cardiovascular disease. XGBoost is well-known for its effectiveness and regularization strategies that stop overfitting. It works especially well with big and complicated datasets. The model supports extensive hyperparameter tuning and feature selection, enabling the optimization of the feature set or the introduction of new feature engineering processes.

2.3 Evaluation Indicators

Four indicators are leveraged to measure model's performance.

Accuracy (ACC) measures how well the model performs overall by dividing the number of correctly predicted cases by the total number of instances.

Precision (PRE) determines the ratio of true positives to total expected positives (true positives + false positives) to assess the accuracy of the positive predictions.

Recall (REC) measures how well the model detects all actual positive events by dividing the number of true positives by the total number of false negatives and true positives.

F1-score is the harmonic mean of recall and precision. It provides a balanced metric where trade-offs between recall and precision are necessary, particularly when class distributions are unbalanced.

3 EXPERIMENT AND RESULTS

3.1 Experimental Details

The experiments were conducted on a machine with the following configuration: The operating system was Windows 11, driven by a Core i5-11400H 11th generation Intel CPU. The system was equipped with an NVIDIA GeForce RTX 3050 GPU featuring 4GB of VRAM. This setup ensured efficient processing and performance during the experiments.

The code was implemented using Python 3.9.13. The key libraries used in the experiments included Scikit-learn (1.0.2) for machine learning models, XGBoost (2.0.3) for gradient boosting, Pandas (1.4.4) for data manipulation.

3.2 Hyperparameters

To guarantee reproducible findings, the dataset was split into80% training and 20% testing. A random seed of 40 was used. The max number of iterations for the Logistic Regression model, which served as a benchmark for comparison, was set at 1000. The Decision Tree had two fixed parameters: the maximum depth was four, and the smallest sample size required to divide a node was five thousand. At least of 100 samples were required to divide a node in the Random Forest model, which included 100 Decision Trees with 10 as the largest. Lastly, the XGBoost model used a total of 2000 boosted trees, a learning rate of 0.01, verbosity level set to 1 for outputting messages, a largest tree depth of 4, a smallest weight of 1 needed for a child node, a training sample ratio of 0.7, a column sampling ratio of 0.7, and early stopping if there was no improvement in performance for 20 consecutive rounds.

3.3 Performance Comparison

In this experiment, various techniques were employed to enhance model performance, such as encoding categorical variables, creating additional polynomial features, filtering out features with minimal variability, and selecting the most relevant features. Data was also standardized to ensure uniformity across features, and an extensive search was conducted to identify the best hyperparameters. The results provide a summary of the evaluation measures that were used to determine how effective various strategies were. Table 1 computes and presents them.

Table 1: Comparison of several models' performances without the use of performance-enhancing methods.

	ACC	PRE	REC	F1
Logistic Regression	.6989	.7084	.6687	.6880
Decision Tree	.6350	.6315	.6355	.6335
Random Forest	.7201	.7252	.7024	.7136
XGBoost	.7297	.7410	.7002	.7200

Based on the previously indicated specific methodologies, Table 2 presents the model scores.

Table 2: Comparison of several models' performances with the use of performance-enhancing methods.

	ACC	PRE	REC	F1
Logistic Regression	.7131	.7185	.6936	.7059
Decision Tree	.7318	.7488	.6916	.7191
Random Forest	.7372	.7502	.7056	.7272
XGBoost	.7367	.7507	.7031	.7261

Taking the accuracy rate as the principal index, the Random Forest Model's performance is the best, and the Logistic Regression Model's performance is worst. In general, the Logistic Regression Model's performance is worse than the other three models. Meanwhile, the other three models performed similarly and have got close scores, only three decimal places apart. Additionally, all of scores of four models increased to varying degrees, especially the Decision Tree Model, almost increased 10%. It seemed like that the influence of data preprocessing and the hyperparameter settings is less than choosing different models.

4 DISCUSSIONS

4.1 Model Performance Analysis

Based on the experimental results, different models exhibit significant performance differences before and after preprocessing, revealing their respective strengths and weaknesses. Prior to preprocessing, the Logistic Regression Model's accuracy was 0.6989 and its F1 score was 0.6880. Its accuracy increased to 0.7131 and its F1-score to 0.7059 following preprocessing. This suggests that linear models perform much better when feature engineering is applied. However, due to its linear nature, Logistic Regression struggles to capture complex nonlinear relationships, leading to its performance being inferior to more sophisticated models. The decision tree model's accuracy was 0.6350 and its F1 score was 0.6335 when the data was not treated, clearly indicating an overfitting bias. After preprocessing, the Decision Tree's accuracy rose to 0.7318 and its F1-score to 0.7191, demonstrating that feature selection and standardization effectively mitigated overfitting issues. Using an ensemble of Decision Trees to minimize overfitting, the accuracy was 0.7201 and the F1-score was 0.7136 produced by the Random Forest on the raw data; these increased to 0.7372 and 0.7272, respectively, following preprocessing, demonstrating the model's resilience and capacity for generalization. Using the raw dataset, the accuracy was 0.7297 and the F1-score was 0.7200. And it improved to 0.7367 and 0.7261 after preprocessing — XGBoost again demonstrated remarkable performance. In this experiment, the Random Forest model significantly outperformed XGBoost, despite the latter using gradient boosting and regularization techniques to handle high-dimensional and nonlinear data successfully. This is mainly because Random Forest's ensemble method, which lowers overfitting and better captures complicated patterns by averaging several trees, performs better. Overall, data preprocessing significantly improved the accuracy across various approaches, with Random Forest demonstrating the highest robustness and accuracy, indicating that in this analysis, it is the best accurate model for predicting cardiovascular disease.

4.2 Limitations and Future Works

Although data preprocessing improved the performance of the models, there's still potential for additional optimization. Feature selection and engineering strategies could be more comprehensive, particularly by incorporating domain knowledge or exploring automated feature generation techniques to enhance model accuracy and generalization. Furthermore, in order to handle more complicated data patterns, future research may take into account integrating deep learning models, like neural networks, into the trials instead of solely depending on classic machine learning models. Moreover, optimizing hyperparameter search methods and using techniques like cross-validation could further improve model robustness and predictive accuracy. In practical applications, increasing the data volume, refining the feature set, and employing multi-model fusion strategies could potentially provide more reliable and accurate predictions for cardiovascular disease.

5 CONCLUSIONS

In this study, the author evaluated the performance of several machine learning models for predicting cardiovascular disease, focusing on four models. The analysis incorporated extensive preprocessing

techniques, including One-Hot Encoding, Polynomial Feature Expansion, feature selection, and data standardization, to enhance model performance and robustness.

The experimental results revealed that the greatest accuracy and F1-score were attained by Random Forest, outperforming other models. Specifically, after preprocessing, Random Forest's accuracy reached 0.7372 and its F1-score improved to 0.7272, showcasing its strong generalization ability and exceptional handling of complicated data patterns. XGBoost fared much better, with excellent F1-score and accuracy but lagging slightly behind Random Forest. The improvements observed in all models, particularly the significant gains for Decision Tree and Random Forest, emphasize how important feature engineering and preprocessing are to improving the performance of the model.

In order to handle even more complicated data patterns, future work should investigate more thoroughly how medical domain knowledge may be integrated into feature selection and take into account cutting-edge methods like deep learning models. Moreover, utilizing multi-model fusion techniques and expanding the dataset can further increase anticipated dependability and accuracy. This study highlights the significance of preprocessing the raw data and choosing a suitable model to obtain the highest predictive performance for cardiovascular disease prognosis.

REFERENCES

Biau, G., & Scornet, E. 2016. A Random Forest guided tour. *Test*, *25*, 197-227.

Celermajer, D. S., Chow, C. K., Marijon, E., Anstey, N. M., & Woo, K. S. 2012. Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection. *Journal of the American College of Cardiology*, *60*(14), 1207-1216.

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.

Hagan, R., Gillan, C. J., & Mallett, F. 2021. Comparison of machine learning methods for the classification of cardiovascular disease. *Informatics in Medicine Unlocked*, *24*, 100606.

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

Knutsson, A., & Bøggild, H. 2000. Shiftwork and cardiovascular disease: review of disease mechanisms. *Reviews on environmental health*, *15*(4), 359-372.

Kononenko, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109.

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

Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. 2020. Artificial intelligence, machine learning, and cardiovascular disease. *Clinical Medicine Insights: Cardiology*, *14*, 1179546820927404.

Priyanka, & Kumar, D. 2020. Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246-269.

Svetlana, U., 2018. Cardiovascular Disease dataset. URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. Last Accessed: 2024/09/18