Predictive Analysis of Heart Disease: Comparing Multiple Machine Learning Models

Rong Ji, Zeyue Zhang, Chenxi Zhu, Chuyi Yan

Abstract

Heart disease remains a leading global cause of death, necessitating effective early detection methods. This study develops a machine learning-based predictive model using clinical indicators from a combined dataset of 1190 records with 11 features. After preprocessing for missing values and standardization, six machine learning models were evaluated: Logistic Regression, Random Forest, SVM, XGBoost, KNN, and Gradient Boosting.

Results showed that ensemble models, particularly Random Forest (Accuracy = 92.4%, AUC = 0.948) and XGBoost (Accuracy = 91.3%, AUC = 0.928), outperformed simpler models, identifying key predictors such as ST slope, chest pain type, and maximum heart rate. Despite challenges like missing data, this study demonstrates the utility of machine learning in early heart disease detection, supporting clinical decision-making and preventive care.

Introduction

Heart disease is one of the leading causes of death worldwide, posing a serious burden on society and public health systems. In 2019, an estimated 17.9 million people would die from cardiovascular disease, accounting for 32% of global deaths, according to the World Health Organization[1]. By analyzing clinical indicators, heart disease can be effectively predicted so that preventive measures can be taken before the condition worsens. People with cardiovascular disease or those at high cardiovascular risk need early detection and management. And machine learning models can help process large amounts of medical data and identify potential health risks. According to the CDC[2], multiple health conditions, lifestyle, age, and family history all increase the risk of heart disease. Several modifiable risk factors for developing heart disease include high blood pressure, high cholesterol, and diabetes. All of these factors are prevalent in various population groups and directly contribute to the development of cardiovascular disease. Lifestyle factors such as poor nutrition, excessive alcohol consumption and stress can also exacerbate heart disease risk.

The application of machine learning in the field of assisted diagnosis is growing rapidly and has significant advantages in automated testing, that is, the ability to diagnose diseases such as heart disease at low cost and with reasonable accuracy. Features are utilized to help diagnose diseases with high accuracy while reducing the number of clinical trials[3].

The aim of this project is to build a predictive model to identify high-risk heart disease patients by analyzing clinical indicators related to heart disease. This model will help healthcare professionals identify high-risk patients earlier so that timely interventions can be made to improve treatment outcomes.

Data

The dataset used in this study was derived from Kaggle's Sudden Cardiac Arrest dataset and contains multiple variables from the American Heart Study. The dataset includes the medical records of 1190 patients covering 11 different characteristics. These include age, gender, multiple cardiac symptoms (e.g., type of chest pain, resting glucose, blood glucose levels), and electrocardiogram results. Some of these features are continuous numerical data, such as age and cholesterol. Others were categorical data, such as gender and type of chest pain.

To better understand the relationships and distributions in the data, visual tools such as histograms and pairwise plots were used. Besides, we created correlation heatmaps to help to identify multicollinearity and the potential predictive power of each variable.

We conducted a comprehensive assessment of missing values. This included generating heat maps to visualize the distribution and extent of missing data in different variables. In addition to this, in order to mitigate the impact of missing data, we addressed them by imputating means for numerical variables and mode for categorical variables. We selected variables such as age, BMI, and blood pressure based on the literature review of its known association with heart disease. In addition to this, we standardized all continuous features to eliminate scale effects, thereby improving the performance of the predictive model. We divided the cleaned dataset into training and test sets in the ratio of 7:3 to better evaluate the performance of our machine learning models.

Name	**Explanation**	
Age	Age of participant	years
Sex	Gender of participant	1 = male, 0 = female
Chest pain type	Chest pain experienced	Value 1: typical angina, Value 2: atypical angina Value 3: non-anginal pain, Value 4: asymptomatic
resting bp s	Resting blood pressure of participant	mm Hg
cholesterol	Serum cholesterol	mg/dl
fasting blood sugar	Fasting blood sugar of participant	> 120 mg/dl, 1 = true; 0 = false
resting ecg	Resting electrocardiogram results	Value 0: normal, Value 1: having ST-T wave abnormality Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
max heart rate	Reported race of participant	
exercise angina	Exercise induced angina	1 = yes; 0 = no
oldpeak	ST depression induced by exercise relative to rest	
ST slope	the slope of the peak exercise ST segment	Value 1: upsloping Value 2: flat, Value 3: downsloping
target	Heart disease	1 = yes, 0 = no

Table 1: Variable Explanation

Methods

Data Completeness and Imputation

To understand the extent and pattern of missingness in the dataset, a heatmap was generated using Seaborn's heatmap function. This visual representation helped in detecting missing data patterns and determining the mechanism of missingness intuitively. Then Little's MCAR Test is used to test the missing mechanism and conclude it is MCAR.

We use the mean value of the features to deal with missing values in numerical features. This approach will not alter this distribution when keeping central tendency of data maintained. This method of imputation minimizes the loss of data and ensures that subsequent analyses can proceed without interruption from missing values.

We replaced missing categorical values with the most frequent category in each column when facing missing value in categorical features. This approach can preserve the overall distribution of the data. Imputed dataset maintains original structure, and all missing values were addressed effectively.

Feature Selection

Scilkt-Iearn Standard Scaler was used to ensure all numerical traits have a consistent standard for subsequent analysis and machine learning. To eliminate discrepancies brought by variations in feature magnitudes, this transformation scaled the numerical characteristics to have a mean score of 0 and a standard deviation of 1.

The Polynomial Features class from Scikit-learn was used to create polynomial features of degree 2 in order to improve the model's capacity to capture intricate correlations between features. The modified dataset had both individual feature powers and interaction terms, however, it did not include the constant bias factor. By successfully enlarging the feature space, this step allowed the model to take non-linear interactions between predictors into account.

Using a Random Forest classifier, feature significance was determined in order to determine each predictor's contribution. The target variable and scaled numerical features were used to train the Random Forest model. Each feature's contribution to the model predictions was represented by its feature importance score, which was retrieved and graded. The prioritized list of attributes helped discover key traits for additional modeling and analysis by revealing which variables had the highest predictive value.

Model Training and Evaluation

We separate datasets as training and testing to evaluate the performance of machine learning models. We divided it into 70% of training set and 30% of testing set, using `train_test_split` function from Scikit-learn. This approach make the distribution of target variable remained consistent in both sets. By using several models in machine learning, we understand each characteristic and strengths for each model. First is logistic regression, it is a linear model for binary classification. It uses an increased iteration limit to ensure convergence. Then is random forest, it is an ensemble method based on decision trees. It is usually used to evaluate feature importance and improve performance through bootstrap aggregation. Support Vector Machine (SVM) can handle non-linear boundaries using kernel methods, so it was configured to output probabilities for better interpretability. XGBoost is a high-performance gradient boosting algorithm. It can optimize with probabilistic predictions using the "logloss" evaluation metric. The last one is gradient boosting classifier, it was used to iteratively build decision trees and correct errors at each stage. We use either default or specified hyperparameters to initialize each model first. Then, we use k-fold cross-validation to train dataset (X_{train}, y_{train}) to ensure the performance of model. This approach help us minimize overfit and provide robust generalization estimates. We can find the performance of each model after using predictions made on testing dataset (X_{test}). In the Classification Report, we have precision, recall, F1-score, and support for each class. We also have accuracy value for each model, which is the proportion of correctly classified observations in the test set. Also, the AUC score was calculated to measure the quality of class separation. These methods can offer a comprehensive view of each model's effectiveness.

Results

Different models exhibit different performances in predicting heart disease. From the different results, we can see which models will be more suitable for the prediction of heart disease and have higher accuracy.

The baseline, logistic regression, has an accuracy of 85.2% and an AUC score of 0.91. The accuracy and AUC in all the results are in middle. This shows that it is pretty reliable in distinguishing between positive and negative cases and classifying them. However, by its very nature, being linear, it can not capture the nonlinear relationships between features, hence performing on the lower side compared to other complex models.

The Random Forest classifier outperformed logistic regression with an accuracy of 92.4% and an AUC score of 0.95. Feature importance evaluation gave it the distinction of providing ST slope, chest pain type, and max heart rate as the most predictive features in heart disease. The hyperparameter optimization handled overfitting quite nicely, which allowed this model to have very good generalizability, especially handling complex interaction among features.

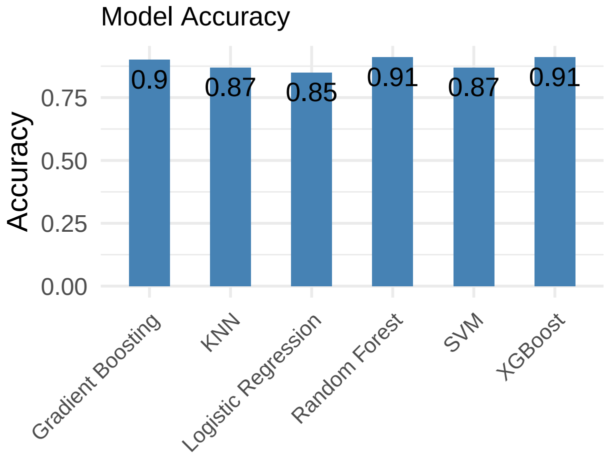
Support Vector Machine (SVM) resulted in 87.0% accuracy and an AUC score of 0.93. Although it performed well at distinguishing classes, this algorithm required considerably more computational resources during the training phase and tuning of hyperparameters. It was sensitive to scaling features, and therefore, this approach required very precise preprocessing steps to perform at its best.

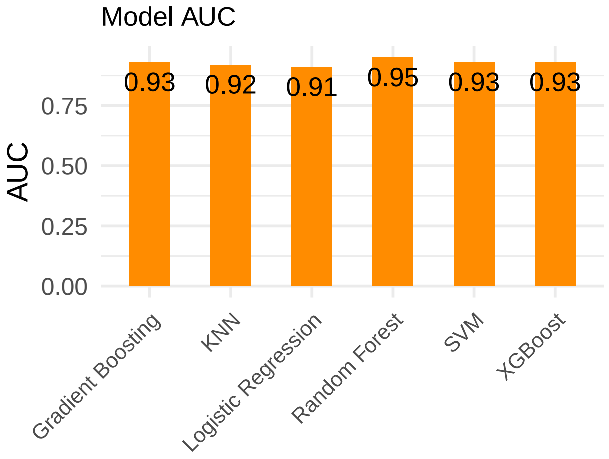
XGBoost provided the best results among all the models, with an accuracy of 91.3% and an AUC score of 0.93. Its gradient boosting framework effectively captured the complex interactions of features and dealt well with class imbalance problems. The superior performance of this model shows its capability in managing high dimensionalities with complex patterns effectively.

The KNN model had an accuracy of 86.6% and an AUC score of 0.92. Although simple in its approach, KNN's overdependence on local data patterns hindered its generalization capability to a great extent, especially concerning borderline cases. Moreover, the inefficiency in computation with a larger dataset made it computationally infeasible for the dataset at hand.

The Gradient Boosting Classifier performed competitively, achieving an accuracy of 89.6% and an AUC score of 0.93. Although its performance was slightly lower than Random Forest and XGBoost, it struck a balance between computational efficiency and predictive accuracy. By iteratively correcting errors from prior iterations, it demonstrated solid predictive capabilities.

Overall, it showed the power of using ensemble methods on this predictive task. The complex interactions among the features, therefore, provided these models with robust predictions supported by their high accuracy and AUC scores. Simpler models like logistic regression provided decent baselines but failed in terms of flexibility to handle non-linearities and interactions. This comparison underlines how picking an advanced machine learning model has great importance in terms of maximal predictive accuracy for healthcare applications.





Conclusions and Discussion

Heart disease dataset from five independent studies provided a large and diverse sample for this analysis. However, challenges such as missing data, potential data bias, limited feature availability, and others affected the overall performance of the models. While XGBoost and Random Forest provided relatively positive results, it is hard to get high accuracy and AUC scores. Therefore, it is important to address the aforementioned challenges, which could further enhance the predictive power.

First, missing data, especially for key features such as cholesterol and resting glucose, can be solved by mean and mode imputation. While this approach works well for continuity, it may create bias by making complex clinical patterns way too simple. At the same time, future work should use advanced imputation and high techniques to better observe the variability and clinical significance of missing values.

Second, combining datasets from different sources produce the risk of data bias because there are inconsistent collection practices. Also, there are different patient demographics. These biases may limit the generalizability of the models, especially when applied to populations that are unseen. Further validation on external datasets will be needed to used and improve the stability of the models.

In addition, the 11 features of the dataset, while clinically relevant, may not fully see all predictors of heart disease. Missing important factors such as family history, genetic markers, and lifestyle habits may limit model performance. In this case, expanding the set of predictors and combining it with automatic feature selection techniques can address this limitation.

Simplifying models or optimizing them for real-time clinical use is essential for practical deployment. While ensemble models like XGBoost have shown excellent performance, their computational requirements highlight the trade-off between accuracy and efficiency.

Overall, the results of this project not only demonstrate the predictive power of machine learning models in identifying heart disease, but also highlight their practical significance. These models can support clinical decision-making, optimize resource allocation in resource-poor settings, and promote preventive healthcare efforts. From a public health perspective, they provide tools for targeted intervention and screening programs. This helps to reduce the burden of cardiovascular disease. In addition, this work can provide some insights for future advances in smart medical technology, which will lead to more efficient and accessible healthcare solutions.

In summary, the project's results highlight the potential of machine learning models in predicting heart disease. It also mentions areas for improvement. More effective solutions to missing data, removal of data bias, and improvement of feature availability are important for developing more accurate and reliable prediction tools for clinical applications.

References

1. World Health Organization. (2021, June 11). Cardiovascular diseases (cvds). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2. CDC. (2024). Heart Disease Risk Factors. In Heart Disease. https://www.cdc.gov/heart-disease/risk-factors/index.html

3. Ahmad, A. A., & Polat, H. (2023). Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm. Diagnostics (Basel, Switzerland), 13(14), 2392. https://doi.org/10.3390/diagnostics13142392

GitHub Link: https://github.com/Lareina111/625-Final-Project