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.It is curated by combining 5 independent heart disease

datasets.The five datasets used for its curation are: Cleveland, Hungarian, Switzerland, Long Beach VA and Statlog (Heart) Data Set. 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 characteristics are continuous numeric data, such as age and cholesterol, while others are categorical data, such as sex and chest pain type.

To better understand the relationships and distributions in the data, visual tools such as histograms and pairwise plots were used to understand the distribution of the variables and the relationships between them. In addition, we created correlation heatmaps to examine the relationships between the variables, which helped 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, helping us to identify patterns of missing data and informing subsequent imputation strategies. To mitigate the effects of missing data, we addressed the issue of missing values by interpolating means for numerical variables and modes for categorical variables. We selected variables such as age, BMI, and blood pressure based on their prevalence in the survey cycle and 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.

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.

Missing values in the numerical features were imputed using the mean of the respective feature. This approach ensures that the central tendency of the data is maintained without significantly altering its distribution. This method of imputation minimizes the loss of data while ensuring that subsequent analyses can proceed without interruption from missing values.

To handle missing data in categorical variables, the process began by identifying all categorical features in the dataset. For columns with partially missing values, the missing entries were replaced with the most frequent category (mode) in each column to preserve the overall distribution of the data. Since there were no columns with entirely missing values, all features were retained for analysis. The final imputed dataset

maintained its original structure, ensuring that missing categorical values were addressed effectively without introducing bias or distorting the integrity of the data.

Feature Selection

To ensure that all numerical characteristics had a consistent scale for subsequent analysis and machine learning, standardization was performed using the Scikit-learn StandardScaler. This transformation scaled the numerical features to have a mean of 0 and a standard deviation of 1, eliminating disparities caused by differences in feature magnitudes.

To enhance the model's ability to capture complex relationships between features, polynomial features of degree 2 were generated using the PolynomialFeatures class from Scikit-learn. Both individual feature powers and interaction terms were included in the transformed dataset, while the constant bias term was excluded. This step effectively expanded the feature space, enabling the model to account for non-linear interactions among predictors.

To identify the contribution of each predictors, feature importance was calculated based on a Random Forest classifier. The Random Forest model was trained on the scaled numerical features (X) and the target variable (y). Feature importance scores that represent the contribution of each feature to the model predictions were extracted and ranked. The ranked list of characteristics provided insight into which variables had the strongest predictive power, helping identify important characteristics for further modeling and analysis.

Model Training and Evaluation

The preprocessed dataset was divided into training and testing subsets to evaluate the performance of various machine learning models. A stratified split was applied using the `train_test_split` function from Scikit-learn, with 70% of the data allocated to the training set and 30% to the testing set. This ensured that the distribution of the target variable (y) remained consistent across both sets while reserving unseen data for model evaluation.

Several machine learning models were employed for the classification task, each with distinct characteristics and strengths. Logistic Regression, a linear model for binary classification, was utilized with an increased iteration limit to ensure convergence. Random Forest, an ensemble method based on decision trees, was applied to evaluate feature importance and improve robustness through bootstrap aggregation. Support Vector Machine (SVM), with its ability to handle non-linear boundaries using kernel methods, was configured to output probabilities for better interpretability. XGBoost, a high-performance gradient boosting algorithm, was optimized with probabilistic predictions using the "logloss" evaluation metric. K-Nearest Neighbors (KNN), a non-parametric method, classified observations based on distances in the standardized feature space. Finally, Gradient Boosting Classifier, a boosting ensemble technique, was used to iteratively build decision trees, correcting errors at each stage. Each model was initialized with either default or specified hyperparameters and trained on the training dataset (X_{train}, y_{train}) using k-fold cross-validation to ensure reliable performance evaluation. This approach ensured that each model was rigorously validated while minimizing overfitting and providing robust generalization estimates.

The performance of each model was assessed using predictions made on the testing dataset (X_{test}). Key evaluation metrics included the Classification Report, which provided precision, recall, F1-score, and support for each class. Accuracy was computed as the proportion of correctly classified observations in the test set. Additionally, for models supporting probabilistic outputs, the AUC score was calculated to measure the quality of class separation. These metrics offered a comprehensive view of the models' effectiveness in handling the classification task.

Results

The performance of the machine learning models applied to predict heart disease demonstrated significant differences in accuracy, precision, recall, and AUC metrics, reflecting varying strengths and limitations.

Logistic regression, used as a baseline, achieved an accuracy of 85.2% and an AUC score of 90.9%. Its precision and recall were balanced across the target classes, indicating that it was able to classify both positive and negative cases reliably. However, due to its linear nature, it struggled to capture non-linear relationships among the features, resulting in lower performance compared to more complex models.

The Random Forest classifier outperformed logistic regression, achieving an accuracy of 92.4% and an AUC score of 94.8%. Its ability to evaluate feature importance identified ST slope, chest pain type, and max heart rate as the most influential predictors of heart disease. Through hyperparameter optimization, the model demonstrated excellent generalizability by reducing overfitting and handling complex interactions among features effectively.

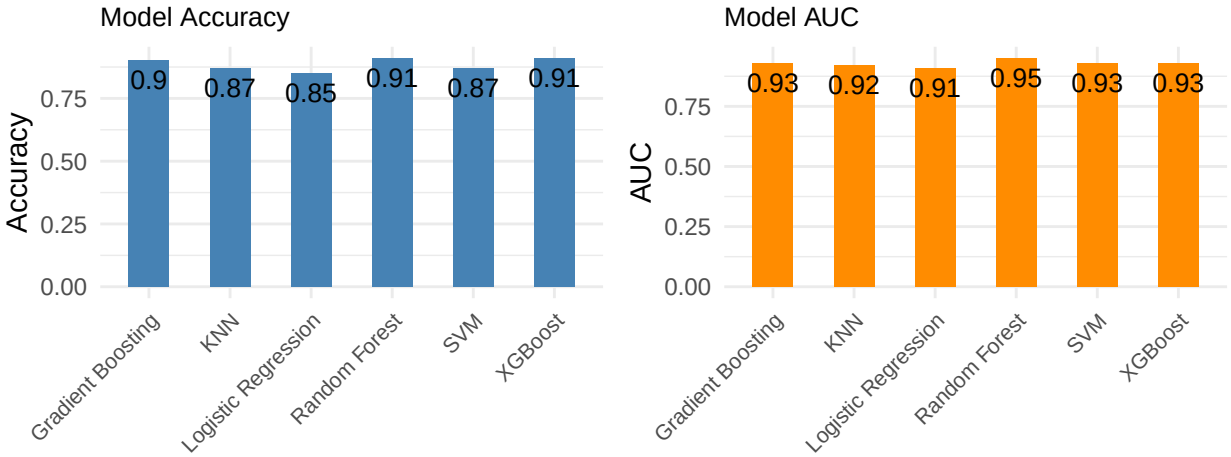
Support Vector Machine (SVM) delivered an accuracy of 87.0% and an AUC score of 93.2%. While it performed well in separating the classes, the computational cost of training the model and tuning hyperparameters was notably higher. Its sensitivity to feature scaling required precise preprocessing steps to ensure optimal performance.

XGBoost provided the best results among all the models, with an accuracy of 91.3% and an AUC score of 92.8%. Its gradient boosting framework was effective in capturing intricate feature interactions and addressing class imbalance issues. This model's superior performance underscores its ability to handle high-dimensional data and complex patterns efficiently.

The K-Nearest Neighbors (KNN) model achieved an accuracy of 86.6% and an AUC score of 91.6%. Despite its simplicity, KNN's reliance on local data patterns limited its ability to generalize effectively, particularly for borderline cases. Additionally, its computational inefficiency when dealing with larger datasets highlighted its limitations in this context.

The Gradient Boosting Classifier performed competitively, achieving an accuracy of 89.6% and an AUC score of 93.0%. 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, the results highlighted the superior performance of ensemble methods like XGBoost and Random Forest in this predictive task. These models leveraged the complex interactions among features, providing robust predictions with high accuracy and AUC scores. In contrast, simpler models like logistic regression offered reliable baselines but lacked the flexibility to handle non-linearities and interactions effectively. The comparison of results underscores the importance of selecting advanced machine learning models to maximize predictive accuracy in healthcare applications.



Conclusions and Discussion

The combined heart disease dataset, sourced from five independent studies, provided a large and diverse sample for this analysis. However, challenges such as missing data, potential data bias, and limited feature availability impacted the models' overall performance. While XGBoost and Random Forest delivered the relatively best results, achieving high accuracy and AUC scores, addressing these challenges could further enhance predictive capabilities.

Missing data, particularly in key features like cholesterol and resting blood sugar, was handled through mean and mode imputation. While effective for continuity, this method may have introduced bias by oversimplifying complex clinical patterns. Future work should adopt advanced imputation techniques to better preserve the variability and clinical significance of missing values.

Combining datasets from different sources introduced the risk of data bias due to inconsistent collection practices and varying patient demographics. These biases may limit the generalizability of the models, particularly when applied to unseen populations. Further validation on external datasets is necessary to assess and improve model robustness.

The dataset's 11 features, though clinically relevant, may not fully capture all predictors of heart disease. Important factors such as family history, genetic markers, and lifestyle habits were absent, potentially constraining model performance. Expanding the predictor set and incorporating automated feature selection techniques could address this limitation.

While ensemble models like XGBoost demonstrated superior performance, their computational demands highlight the trade-off between accuracy and efficiency. Simplifying these models or optimizing them for real-time clinical use will be essential for practical deployment.

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 under-resourced settings, and facilitate preventive healthcare efforts. From a public health perspective, they offer tools for targeted interventions and screening programs, ultimately contributing to the reduction of cardiovascular disease burden. Furthermore, this work could provide some insights for future advancements in intelligent healthcare technologies, leading to more personalized, efficient, and accessible healthcare solutions.

In summary, the results underscore the potential of machine learning models in predicting heart disease but also highlight areas for improvement. Addressing missing data more effectively, mitigating data bias, and enhancing feature availability will be critical for developing more accurate and reliable predictive 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