Analysis of Supervised Machine Learning Models for Heart Disease Prediction

Eric Ohemeng

2024-04-30

Introduction

Cardiovascular diseases (CVDs) remain the foremost cause of death globally, accounting for an estimated 17.9 million lives annually, as reported by the World Health Organization. Heart disease, a significant contributor within the spectrum of CVDs, often goes undetected until it is too late, making early identification and timely intervention critical (Sierra-Galan et al., 2022)

At the heart of this project lies the question: "How well can the presence of heart disease in individuals be predicted based on their clinical and physiological data?" The pursuit of this question is driven by the potential to impact the early detection and treatment of heart disease. Identifying at-risk individuals through predictive modeling can facilitate earlier medical interventions such as lifestyle modification, or further diagnostic testing, which could be lifesaving.

This study harnesses data from the "Heart Disease UCI" dataset, downloaded from Kaggle. Variables such as age, sex, cholesterol levels, and blood pressure readings have been selected due to their proven association with heart health. These factors are not only pivotal in understanding an individual's risk profile but also offer actionable insights. For instance, a prediction model that flags high risk for heart disease can lead to preventive counseling, nutritional advice, and proactive health screenings, thereby mitigating risk before the onset of severe complications.

Exploratory Data Analysis (EDA) and Variable Description

The exploratory data analysis (EDA) and data processing for the Heart Disease UCI dataset laid a solid foundation for predictive modeling, offering insights into key variables and their relationships. EDA uncovered vital distributions and correlations, highlighting factors such as age, resting blood pressure, cholesterol levels, and other indicators directly linked to heart disease. The data processing phase involved cleaning, handling missing observations, etc. to prepare the dataset optimally for machine learning. The processed data with 1018 observations across 14 variables, was then split into training (80 percent) and testing (20 percent) sets, maintaining a appropriate distribution of the binary outcome variable. The subsequent modeling phase compared various algorithms, including KNN and XGBoost, etc.

Table 1: Variable Description

Variable Name	Data Type	Description
Age	Numerical	Age of the patient in years.
Sex	Categorical	Patient's sex: 0 = Female, 1 = Male.
chest_pain_type	Categorical	Type of chest pain experienced by the patient: 1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic.
resting_blood_pressure	Numerical	Resting blood pressure (in mm Hg on admission to the hospital).
cholesterol	Numerical	Serum cholesterol in mg/dl.
fasting_blood_sugar	Categorical	Fasting blood sugar > 120 mg/dl: 0 = No, 1 = Yes.
resting_electrocardiogram	Categorical	Resting electrocardiographic results: 0 = Normal, 1 = ST-T wave abnormality, 2 = Left ventricular hypertrophy.
max_heart_rate_achieved	Numerical	Maximum heart rate achieved during the stress test.
exercise_induced_angina	Categorical	Exercise-induced chest pain: 0 = No, 1 = Yes.
st_depression	Numerical	ST depression induced by exercise relative to rest.
st_slope	Categorical	The slope of the peak exercise ST segment: 1 = Upsloping, 2 = Flat, 3 = Downsloping.
num_major_vessels	Numerical	Number of major vessels (0-4) colored by fluoroscopy.
thalassemia	Categorical	Thalassemia: 1 = Normal, 2 = Fixed defect, 3 = Reversible defect.
Target (Dependent Variable)	Categorical	Presence of heart disease: 0 = Absence, 1 = Presence.

Participants' ages range from 29 to 77, with a median of 56, reflecting the relevance of age as a risk factor. The dataset consists of 309 females and 709 males, highlighting a gender imbalance. Critical variables, such as resting blood pressure and cholesterol, show wide ranges, emphasizing the varied risk profiles for heart disease. ECG abnormalities are prevalent, with 509 participants displaying ST-T wave abnormalities, indicating diverse heart health statuses. The target variable shows a balanced distribution of 495 individuals diagnosed with heart disease and 523 without. The dataset's broad coverage of key health metrics, ranging from exercise-induced angina to ST depression and major vessel counts, provides foundation for exploring predictive models and developing diagnostic strategies for heart disease.

A summary of some of the variables has been presented in the histograms below.

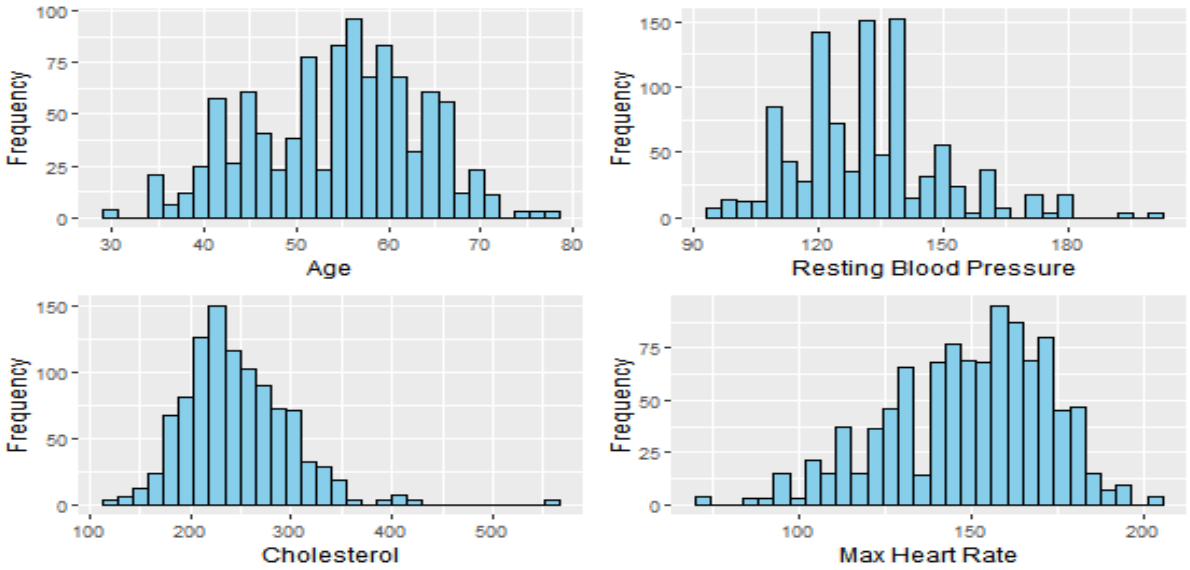

Figure 1: Numerical variables

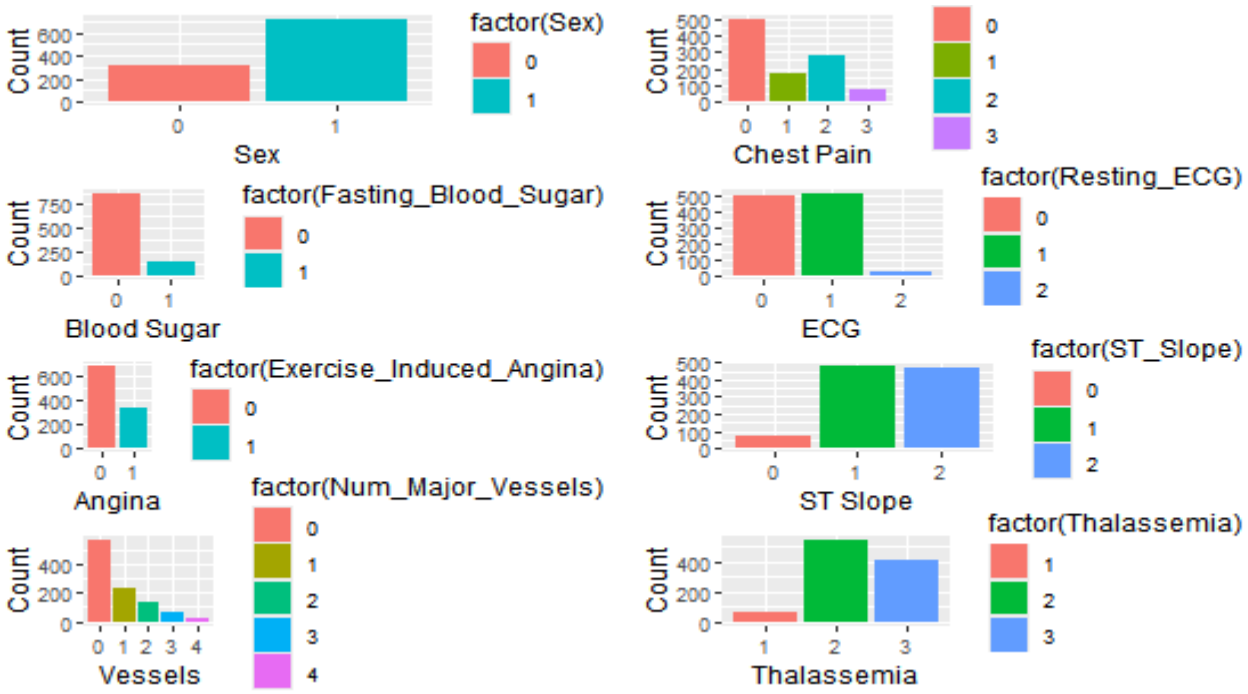

Figure2: Categorical variables

The correlation matrix heatmap below also provides a comprehensive view of the relationships between variables in the Heart Disease UCI dataset. Blue and red shades indicate the degree and direction of these correlations, with blue representing positive correlations and red indicating negative correlations. For instance, a positive correlation between age and ST depression suggests that older individuals tend to have higher ST depression levels. Conversely, a negative correlation between max heart rate and ST depression indicates that individuals with higher maximum heart rates tend to have lower ST depression levels. These relationships reveal key

risk factors for heart disease, including resting blood pressure, cholesterol levels, and ECG results, guiding predictive modeling and clinical interventions. The nuanced understanding from this heatmap informs models in predicting heart disease more accurately, offering a balanced approach to treatment and management.

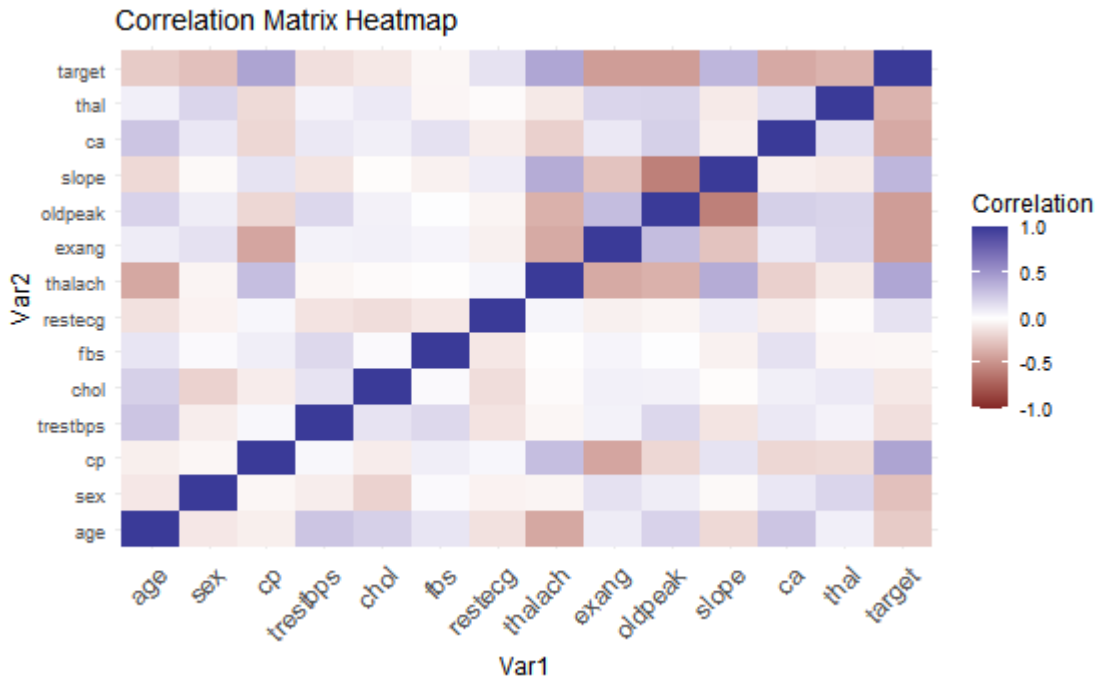

Figure3: Correlation matrix of variables

Model Selection Process

Several models were tried with different tuning parmeters to handle the binary classification task effectively. Logistic regression transformed the response into a logit function, while Lasso regression applied L1 regularization with an alpha of 1 and a range of lambda values, balancing simplicity, and accuracy. Decision trees incorporated a cp parameter, reducing overfitting by limiting splits. Random Forest explored tree splits via mtry configurations, balancing bias and variance, while KNN tuned its k parameter for accurate local classifications. XGBoost combined various parameters to control tree iterations, learning rates, and maximum depths, optimizing each stage. Lastly, the Super Learner ensemble combined individual learners like SL.mean, SL.glmnet, and SL.ranger via non-negative least squares, with 5-fold cross-validation for balanced weighting.

Table 2: Tuning Parameters for Various Models

Model	Tuning Parameters
Logistic Regression	family = 'binomial'
Lasso Regression	alpha = 1, lambda = 10^seq (-3, 3, 0.5)
Decision Tree	cp = 0.001
Random Forest	mtry = (2, 4, 6)
Bagging	ntree = 100, maxdepth = 30, cp = 0.001
KNN	k = (3, 5, 7, 9, 11)
XGBoost	nrounds = (100, 200, 500), eta = (0.1, 0.01), max_depth = (3, 5, 7), gamma = 0, min_child_weight = 1
Super Learner	SL.mean, SL.glmnet, SL.ranger

Model Evaluation

In evaluating the models, various metrics were utilized to assess their effectiveness in accurately identifying individuals with the condition. Accuracy, recall, precision, and F1 score are key parameters used to evaluate the performance of the classifiers, as outlined in Table 3. Accuracy measures the overall correctness of predictions, indicating the proportion of correctly classified cases out of the total. Precision focuses on the accuracy of positive predictions, ensuring that those labeled as having heart disease are indeed affected, thus minimizing false alarms. Recall assesses the model's ability to detect all positive cases, minimizing the risk of missing individuals with heart disease. The F1 score balances precision and recall, offering a comprehensive evaluation of the model's performance. Together, these metrics provide valuable insights into the model's effectiveness in identifying and distinguishing individuals with heart disease, crucial for timely interventions and patient care.

The formula for each metric is written below.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{..... eqn 1}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{.....eqn 2}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{.....eqn 3}$$

$$F-1 = \frac{2*Precision*Recall}{Precision+Recall} \quad \text{.....eqn 4}$$

Below are the models fitted and how each performed per the defined metrics above.

Logistic Regression

Logistic regression is a generalized linear model designed for binary classification. It models the relationship between predictors and the log-odds of the target outcome, specifically the presence or absence of heart disease. The model offers clear coefficients, making it particularly interpretable, and provides probabilistic outputs valuable for decision-making. Logistic regression yielded an accuracy of around 86%, precision of 89%, recall of 82%, and an F1 score of 85%. This balance in metrics indicates that logistic regression effectively identifies key risk factors, providing a reliable diagnosis of heart disease. The balanced precision and recall suggest it is equally adept at avoiding both false positives and false negatives, making it a strong model for healthcare scenarios.

Lasso Regression

Lasso regression extends logistic regression by incorporating L1 regularization, aiding in feature selection by penalizing the coefficients of less relevant predictors. This simplifies the model and prevents overfitting. The model's balanced accuracy was around 85%, precision at 89%, recall at 81%, and an F1 score of 85%. This indicates that Lasso effectively narrows down the most relevant variables for predicting heart disease, striking a balance between model simplicity and predictive power. Its balanced metrics also show its ability to effectively manage false positives and false negatives, making it suitable for healthcare contexts.

Decision Tree

A decision tree creates a hierarchical structure by recursively splitting data based on feature values until terminal nodes represent class labels. This non-linear model offers visual insights into the decision-making process, making it highly interpretable. Its performance metrics include an accuracy of around 83%, precision of 85%, recall of 80%, and an F1 score of 82%. This indicates its practicality in managing various predictive factors for heart disease. The balance between precision and recall shows it can handle both false positives and false negatives, making it a practical choice for visualizing how symptoms and tests lead to diagnoses.

Random Forest

Random Forest constructs a forest of decision trees, each built on random feature subsets. It excels in handling large datasets and multi-dimensional features, reducing overfitting. Its balanced accuracy was around 99%, precision at 100%, recall at 97%, and an F1 score of 98%. This shows its strength in managing complex relationships between predictors and providing accurate heart disease predictions. The high precision and recall show it minimizes both false positives and negatives, making it a powerful and reliable ensemble model for diagnosis.

KNN (K-Nearest Neighbors)

KNN classifies observations based on the majority class among its nearest neighbors. It offers adaptability through adjustable 'k' values, handling noise effectively. Its balanced accuracy was around 68%, precision at 69%, recall at 74%, and an F1 score of 71%. This reflects its simplicity and effectiveness for data with clear clusters, such as patients with similar health profiles. However, its lower balanced accuracy and precision suggest it struggles with predicting true positives and negatives, making it less reliable compared to other models.

XGBoost

XGBoost builds an ensemble of decision trees iteratively, optimizing them to minimize loss functions. This model is efficient at handling complex datasets and non-linear relationships. Its balanced accuracy was around 99%, precision at 100%, recall at 98%, and an F1 score of 99%. This shows its rapid, iterative optimization for accurate heart disease predictions, managing complex relationships between predictors effectively. The high precision and recall show it minimizes both false positives and negatives, making it highly reliable.

Super Learner

The Super Learner ensemble, incorporating models such as SL.mean, SL.glmnet, and SL.ranger, demonstrates robust performance in predicting heart disease. The ensemble heavily relies on SL.ranger, which contributes fully to the ensemble's predictions and yields a notably low risk of 0.0166, emphasizing its reliability. Cross-validation results further showcase its robustness, with an accuracy of 0.9852, reflecting a high proportion of correctly classified observations. The precision of 1 highlights no false positives, while a recall of 0.9697 indicates the model's capacity to correctly identify true positives. The balanced F1 score of 0.9846 emphasizes overall consistency, balancing precision and recall showcasing the ensemble's efficiency in predicting heart disease, making it a reliable model for medical data analysis.

Table 3: Model Performance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8621	0.8901	0.8182	0.8526
Lasso Regression	0.8571	0.8889	0.8081	0.8466
Decision Tree	0.8325	0.8495	0.7980	0.8229
Bagging	0.8424	0.8602	0.8081	0.8415
Random Forest	0.9852	1.0000	0.9697	0.9846
KNN	0.7094	0.6887	0.7374	0.7122
XGBoost	0.9901	1.0000	0.9798	0.9898
Super Learner	0.9852	1.0000	0.9697	0.9846

Conclusion

The predictive models analyzed offer invaluable insights into heart disease diagnosis and management, benefiting both healthcare systems and patients. Ensemble methods like Random Forest and XGBoost provide accuracies approaching 99%, reducing false positives and negatives, allowing for appropriate treatment plans, and minimizing unnecessary interventions. The balanced metrics of models like Logistic and Lasso regressions highlight key risk factors, helping healthcare providers tailor personalized treatment strategies. Additionally, predictive models guide early interventions and lifestyle modifications, preventing disease onset and reducing the overall burden on healthcare systems. These models also direct resource allocation, target areas with higher risk, and shape public health initiatives for effective education. The overall impact includes reduced healthcare costs, technological advancements in medical diagnostics, and improved patient care, making these predictions essential for efficient healthcare delivery.

References

Sierra-Galan, L. M., Bhatia, M., Alberto-Delgado, A. L., Madrazo-Shiordia, J., Salcido, C., Santoyo, B., Martinez, E., & Soto, M. E. (2022). Cardiac Magnetic Resonance in Rheumatology to Detect Cardiac Involvement Since Early and Pre-clinical Stages of the Autoimmune Diseases: A Narrative Review. *Frontiers in cardiovascular medicine*, *9*, 870200. https://doi.org/10.3389/fcvm.2022.870200