Using Machine learning Models to Predict Heart Disease

School of Global Public Health, New York University, New York, NY, USA

Spring 2024, GPH-GU 2338 Machine Learning for Public Health

Helen Liang, Ivy Zhao, Xiaotong Zhao

Problem

The prevalence of heart disease, a general term referring to several types of heart conditions including coronary artery disease (CAD), remains one of the leading causes of death in the United States[1]. Early detection and intervention can significantly decrease the risk of heart disease, promoting better quality of life and reducing heart disease-related mortality. The objective of this project is to perform predictions to identify patients with high risk of developing heart disease and speed up the diagnostic process through utilization of various machine learning models. We aimed to develop a model that can accurately predict the presence of heart disease in patients. We reviewed existing literature on risk factors and the pathophysiology of heart disease to develop predictive models for its early diagnosis.

Data

The Heart Disease dataset is a publicly available dataset (source: UC Irvine Machine Learning Repository) with combined data from the Cleveland, Hungary, Switzerland, and VA Long Beach databases, consisting of a total of 76 variables and 920 observations with each observation containing a patient's medical record. However, this project will only be using the 14 variables that are available for public use, including demographic information, physiological measurements, and patient medical history (see Table 1). These variables come in both categorical (nominal and ordinal) and numerical forms (discrete and continuous) such as gender and age, respectively. The continuous outcome variable, "num", for heart disease diagnosis contains values 0 (angiographic less than 50% diameter narrowing) and 1 (angiographic more than 50% diameter narrowing) and will be transformed into a factor variable with "yes" indicating presence of heart disease and "no" indicating absence of heart disease[2]. Classification techniques will be utilized to accurately predict and categorize diagnosis of heart disease. Our objective is to develop models capable of identifying patients based on these features.

Table 1. Summary of the Heart Disease Datase (n=920), Mean(SD), Median(min, max) are repoeted for continuous variables. Frequencies(%) are reported for categorical variables.

Sociodemographic characteristics	Total (N = 920)
Age	
Mean (SD)	53.51(9.42)
Median (IQR)	54 (47,60)
Sex	
Male	194 (21.09%)
Female	726 (78.91%)
Chest Pain Type	
Typical Angina	46 (5.0%)
Atypical Angina	174 (18.91%)
Non-anginal Pain	204 (22.17%)
Asymptomatic	496 (53.91%0
Resting Blood Pressure (in mmHg)	
Mean (SD)	132.1 (0.97)
Median (IQR)	130(120,140)
Serum Cholesterol (in mg/dl)	
Mean (SD)	199.1 (0.98)
Median (IQR)	223 (175,268)
Fasting Blood Sugar > 120 mg/dl	
FALSE	692 (83.37%)
TRUE	138 (16.63%)
resting electrocardiographic results	
Normal	551 (60.02%)
Having ST-T Wave Abnormality	179 (19. 50%)
Showing Probable	188 (20.48%)
Maximum Heart Race Rchieved	
Mean (SD)	137.5(0.97)
Median (IQR)	140 (120,157)
Exercise Included Angina	
No	528 (61.04%)
Yes	337 (38.96%)
ST Depression Induced by Exercise Relative to Rest	
Mean (SD)	0.88 (0.97)
Median (IQR)	0.5 (0, 1.5)
Slope of the Peak Exercise ST Segment	
Upsloping	203 (33.22%)
Flat	345 (56.46%)
Downsloping	63 (10.31%)
Number of Major Vessels Colored by Fluoroscopy	
0	181 (58.58)
1	67 (21.68%)
2	41 (13.26%)
3	20 (6.47%)
Thalassemia	
Normal	196 (45.16%)
Fixed Defect	46 (10.60%)
Reversible Defect	192 (44.23%)
Diagosis of Heart Disease	
No	411 (44.67%)
Yes	509 (55.33%)

Approach

We developed several predictive machine learning models for heart disease diagnosis using the 14 predictors. During data preprocessing, we checked for missing values (NA) and features "number of major vessels colored" and "thalassemia" are dropped since over half of the observations have missing values in these two features. For remaining features, missing (NA) values in continuous variables with median and categorical variables with mode. We then performed feature selection using forward and backward stepwise by adjusted R^2 and Cp, both selecting the same 10 predictors (the feature "resting blood pressure" is dropped). We split our dataset into 80% training and 20% testing sets for model performance evaluations.

We trained machine learning models for classification task using 5 algorithms including logistic regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest and Gradient Boosting. Then we use k-fold cross-validation to tune hyperparameters to achieve better model performance and avoid overfitting (k = 5 since it is a small dataset).

We used logistic regression model as a baseline model since it is a simple and classic method for classification tasks. Then we incorporated elastic net technique and cross-validation for regularization on the logistic regression model. K-Nearest Neighbors (KNN) was used to estimate the response of a data point by capturing local patterns based on its K-nearest neighbors and hyperparameter K was tuned by cross-validation. Random forest, an ensemble method building upon the base model of decision tree, was used to train model for reduce overfitting of individual trees. Its hyperparameter, mtry, determines the number of variables to randomly sample at each split and was tuned by cross-validation. Gradient boosting is an ensemble method to build decision tree which improves the model by reducing errors by the previous tree iteratively. Gradient boosting was used to train model and hyperparameters including number of trees, interaction depth, shrinkage, and minimum number of observations in nodes are tuned by cross-validation. Support Vector Machines (SVM) is a supervised method to find the best hyperplane that can achieve the best classification performance in the feature space. SVM model was trained using Gaussian kernel for handling non-linear relationships between features and cross-validation was used to tune its regularization hyperparameters, Cost for trade-off between maximizing margin and minimizing error, Sigma for smoothness of decision boundary.

Seven evaluation metrics were used for evaluating model performance, including accuracy, sensitivity, specificity, F1 score, AUC and testing error computed on test dataset. Training errors were computed using training dataset. ROC curves were plotted to visualize discrimination ability of models (see Figure 1).

Evaluation

We used a range of metrics to evaluate model performances, and these include computing the training and testing errors on training and testing sets, computing accuracy, sensitivity, specificity, and F1 score (see Figure 2) using confusion matrix when detecting positive and negative cases separately and computing the area under the ROC (AUCROC) values and visualizing the curves (see Figure 1). Higher AUCROC value (between 0 and 1) indicates a better overall model performance. Training and testing errors helped us understand if the models are overfitting or underfitting while specificity and sensitivity minimizes false positives and false negatives respectively. Accuracy was evaluated using both the accuracy score and the F1 score. An accuracy of 70% or above indicates an acceptable model performance, while an F1 score between 0.8 to 0.9 indicates a well-balanced model with high recall and precision.

We have determined that all our final models used for heart disease diagnosis predictions are models after tuning using k-fold cross-validation except for the logistic regression model, and these include: KNN model, random forest model, gradient boosting, and SVM model. Logistic regression using k-fold cross-validation and elastic net regularization did not yield a very promising result compared to our initial logistic regression model since its accuracy and testing error slightly increased. It might be because our training dataset is too small, especially if the features have varying degree of importance. Tuning for regularization hyperparameters might be not very effective and may cause strong penalty strength, leading to underfitting problem.

Furthermore, we plotted and compared the AUC values and AUCROC curves of the final models and gradient boosting revealed the best AUC value of 0.903. All models achieved promising performances on this binary classification task with high accuracies above 80%. Both random forest and gradient boosting models achieved the highest AUC values, suggesting that they have better discrimination ability on the dataset. It might be because they are all ensemble methods and can handle overfitting problem effectively. We noticed that after tuning the hyperparameter of the gradient boosting model, the training error for the gradient boosting model increased significantly although its accuracy slightly increased. We suspect the reason might be that the gradient boosting model had several hyperparameters that required careful tuning when compared to the random forest model. Since

we have a small training dataset, the hyperparameters might be not tuned effectively and therefore, exhibited poor generalization model performances.

FIGURE 1. Comparison of AUCROC Curves for Final Models.

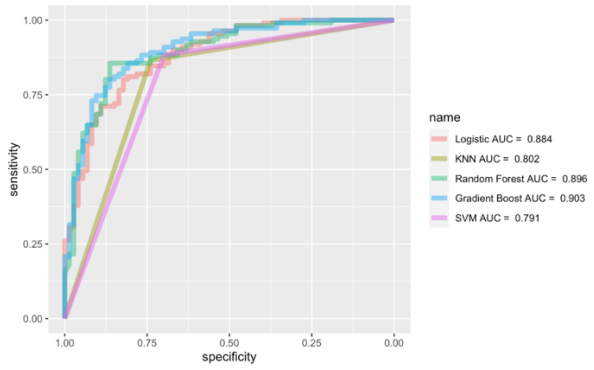
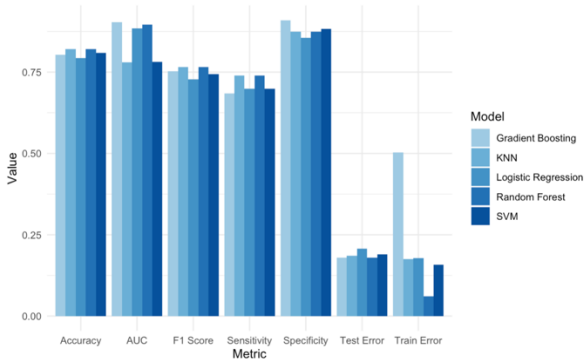


FIGURE 2. Comparison of Accuracy, AUC, F1 Score, Sensitivity, Specificity, Test Error, and Train Error for Final Models.



Conclusion

We trained 5 machine learning models to predict heart disease diagnosis based on 10 features, using algorithms including logistic regression, KNN, SVM, Random Forest and Gradient Boosting. These models all achieve high accuracy of approximately 80%. In combination of all 7 metrics, random forest model outperforms other models, which might be because random forest is an ensemble method that can effectively avoid overfitting in our small dataset.

We did not build Naïve Bayes model mentioned in proposal since it assumes independence between features. However, the existing literature indicates that features in our dataset may have complex relationships, for example, serum cholesterol and blood pressure. Thus, we think Naïve Bayes model is not suitable for our classification task on a small dataset.

In the future, we plan to find larger dataset to train model since the key weakness for our model is that our dataset is small. Furthermore, we plan to incorporate deep learning algorithms such as neural network, which is automatic feature selection to better capture non-linear relationships.

Peer Evaluation

We all work very hard on this project and we agree that everyone should get full points for their work.

Helen Liang: 5/5

Ivy Zhao: 5/5

Xiaotong Zhao: 5/5

References

1. Centers for Disease Control and Prevention. (2023, May 15). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20Disease%20in%20the%20United%20States&text=One%20person%20dies%20every%2033,United%20States%20from%20cardiovascular%20disease.&text=About%20695%2C000%20people%20in%20the,1%20in%20every%205%20deaths.&text=Heart%20disease%20cost%20the%20United,year%20from%202018%20to%202019

2. Janosi Andras, Steinbrunn William, Pfisterer Matthias, and Detrano Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.