Predicting heart disease risk using machine learning: A comparative study of multiple algorithms

Tao Wang

UC Santa Barbara, Santa Barbara, California, 93106, United States of America

taowang@ucsb.edu

Abstract. Heart disease has consistently ranked among the leading causes of morbidity and mortality globally, causing millions of deaths every year, but early diagnosis and medical intervention are considered effective ways to treat heart disease. Therefore, constructing predictive models through data analysis and machine learning algorithms that significantly improve the accuracy of early diagnosis and medical intervention could potentially save millions of lives. Using Kamil Pytlak's dataset from Kaggle, which was originally derived from the 2020 annual CDC survey, this study explores the application of six common machine learning techniques in predicting heart disease. It focuses on data preprocessing, balancing the dataset via undersampling, and feature selection, narrowing down to 8 key risk factors from 17. Among the models—Logistic Regression, LDA, QDA, Boosted Tree, Random Forest, and K Nearest Neighbors—Logistic Regression outperformed others with a 74.6% accuracy and an 82.3% AUC. Despite the challenges in prediction accuracy, the results underline the significant potential of machine learning in early diagnosis and intervention, indicating a promising direction for enhancing public health management strategies against heart disease.

Keywords: Heart disease, Machine learning, Logistic regression, Predictive models

1. Introduction

Heart disease (also known as cardiovascular disease) is currently one of the leading causes of death globally. According to WHO, 17.9 million people died because of heart disease in 2019, accounting for 32% of global deaths [1]. The complexity and fatality of heart disease make it a serious challenge in the medical field. Fortunately, mortality rates of heart disease can be lowered through early warning and intervention [2]. According to the World Health Organization (WHO), most heart diseases can be prevented by changing behavioral risk factors, such as excessive drinking, unhealthy food habits, and smoking [1]. Determining whether an individual is likely to develop heart disease based on their lifestyle habits allows for the proactive reduction of heart disease risk by encouraging high-risk patients to modify their habits.

The analytical capacity of today's algorithms and machine learning models make them promising solutions in society to identify individuals at risk of heart disease in the early stages. With access to large-scale medical data and individual biological information, predictive modeling can be used to assess risk of heart disease based on information such as genetic factors, lifestyle choices, and the presence of other chronic diseases. Early prediction enables patients to preemptively adopt effective health management and prevention measures, thus reducing the incidence rate and mortality of heart disease.

Although current machine learning techniques can predict heart disease, model accuracy is greatly affected by the number and types of risk factors selected [3]. Obtaining the most accurate predictive model remains a significant challenge to biostatisticians today. This research aims to develop some predictive models by using six common machine learning techniques, including Logistic Regression, LDA, QDA, Boosted Tree, Random Forest, and K Nearest Neighbors, and identify the ones with the highest accuracy.

2. Methods

2.1. Data Source

Kamil Pytlak's dataset from Kaggle was utilized for this study. According to Kamil Pytlak, his dataset was derived from the health data of 400,000 adults collected for the 2020 annual CDC survey. The dataset is a big part of the Behavioral Risk Factor Surveillance System (BRFSS), which gathers data on the health status of U.S. residents through annual telephone surveys. Kamil Pytlak cleaned the original CDC survey data and selected the 17 variables most relevant to heart disease to facilitate future machine learning projects on the subject [4]. These risk factors encompass various aspects of an individual's health and lifestyle, providing a holistic approach to modeling and predicting health outcomes, which is why they were all adopted for model construction.

2.2. Data Preprocessing

Although missing values in the original dataset had already been removed, the Kamil Pytlak dataset still required some data processing. To enhance readability of variable names, the clean_names() function in RStudio was used to make variable names more straightforward. Subsequently, class imbalance was addressed. In Kamil Pytlak's dataset of 319,795 observations, a substantial disparity exists: 292,422 observations are related to non-heart disease patients, whereas only 27,373 are related to heart disease patients. The response variable is therefore highly imbalanced and could seriously impact prediction accuracy. The issue was addressed by using undersampling techniques to randomly remove some observations related to non-heart disease patients, thereby maintaining a similar number of observations for non-heart disease patients and heart disease patients. After processing, 27,387 observations are related to non-heart disease patients and 27,373 are related to heart disease patients, facilitating a more balanced response variable. Finally, all non-decimal data type variables were converted to factors.

2.3. Feature Selection

To ensure methodological rigor, the 400,000 data points in the dataset were retested to confirm that there exists a significant correlation between the 17 risk factors selected by Kamil Pytlak and whether the patient has heart disease. This was done through exploratory data analysis, which involved using graphics and summary statistics to explore the relationship between these 17 risk factors and whether the respondent has heart disease. Findings confirmed that all 17 risk factors have a significant impact on whether or not the respondent has heart disease. A generalized linear model was also constructed for this purpose, and the same conclusions were reached.

A random forest model was developed and the varImpPlot() function was used to weigh the relative importance of each risk factor. According to findings, "age_category," "bmi," "gen_health," "sleep_time," "physical_health," "mental_health," "diabetic," "diff_walking" are more important than other risk factors. Including all 17 risk factors in the model may yield good performance on the training data, but it can increase model complexity, potentially leading to overfitting and reduced generalizability to new data. Alternately, selecting only a few risk factors can simplify the model, but this approach may overlook essential information. To obtain a balance between completeness and generalizability, 8 risk factors with the most importance was selected to construct the model. These 8 risk factors are, respectively, "age_category," "bmi," "gen_health," "sleep_time," "physical_health," "mental_health," "diabetic," and "diff_walking."

Table 1. The Attributes of This Study The Attributes of This Study

Attribute	Description	Type
The response variable		
heart_disease	Denotes whether the respondent has ever been diagnosed with heart disease	Booleans
The risk factors		
bmi	Represents the Body Mass Index of the respondent, a measure of body fat based on height and weight.	decimals
physical_health	Represents the number of days in the past 30 days that the respondent experienced poor physical health, including physical illness and injury.	decimals
mental_health	Indicates the number of days in the past 30 days that the respondent experienced poor mental health.	decimals
diff_walking	Denotes whether the respondent has serious difficulty walking or climbing stairs.	Booleans
age_category	Represents the age group of the respondent, with possible answers being '18-24', '25-29', '30-34', '35-39', '40-44', '45-49', '50-54', '55-59', '60-64', '65-69', '70-74', '75-79', '80 or older'.	Strings
diabetes	Denotes whether the respondent has diabetes, with possible answers being 'No', 'No, borderline diabetes', 'Yes', 'Yes (during pregnancy)'.	Strings
gen_health	Represents the respondent's self-assessment of their general health, with possible answers being 'Very good', 'Good', 'Excellent', 'Fair', 'Poor'.	Strings
sleep_time	Indicates the number of hours of sleep the respondent gets in a 24-hour period.	decimals

2.4. Data Split

After feature selection was completed, remaining risk factors were deleted from the dataset. The dataset was split into 2 parts: 80% for training and 20% for testing. The training set has about 43,807 observations, and the testing set has just 10,953 observations. This data splitting method ensures that there would be sufficient data to train the model. The accuracy of our model will not be affected by insufficient training data.

2.5. Model Validation

K-fold cross-validation was used to ensure the model's robustness and evaluate its performance across different subsets of the training data. The 5-fold technique was implemented, and the folds were stratified by the "heart_disease" variable. This approach ensures the distribution of "heart_disease" (often the outcome) remained the same across resamples or, in cross-validation, across folds.

2.6. Logistic regression, LDA, QDA

Logistic regression, LDA, and QDA are three standard machine learning algorithms, often used for binary classification. Logistic regression is a statistical model for classification problems, and it predicts a binary result by using one or more independent variables. Logistic regression has been used in many studies to discover the relationship between disease events and risk factors [5]. It is ideal for analyzing health data and evaluating disease risk factors because its output can be directly interpreted as the probability of disease risk. LDA is a method of finding the best classification boundary between different categories. It is considered a Bayes optimal classifier when the feature distributions in the classes follow a normal distribution with the same covariance matrix [6]. The purpose of LDA is to map the samples from an N-dimensional space to a one-dimensional subspace [7]. QDA is a generalization of LDA. Unlike LDA, QDA does not consider the assumption that the covariance of each class is the same [8]. In QDA, each class can have its own covariance matrix. This can capture more complex relationships between features and classes compared to LDA. However, it requires estimating more parameters, which requires our sample size to be sufficiently large.

These three models were first specified using different functions and engines. For example, a logistic regression model was set for classification by using the logistic_reg() function and the “glm” engine. Then, three workflows were created for different models, and the appropriate recipe introduced. After that, each model workflow was fitted to folded data using the fit_resamples() function, and the collect_metrics() function was employed to identify the model with the highest accuracy and the greatest AUC values. Finally, the workflow of the best model was fitted to the entire training dataset (not to the folds). With this fitted model, predict(), bind_cols(), and accuracy () functions were used to assess model accuracy on the testing data. augment() and roc_auc() functions were also used to calculate the model’s AUC values. Model accuracy and AUC value were all recorded as study results.

2.7. *Random Forest*

Random Forest is an effective machine learning algorithm that works very efficiently for classification. In the setting of random forest, numerous classification and regression trees are built using random selected training datasets and subsets of random predictors for modeling outcomes. These individual trees’ outcomes are aggregated to predict each data point. As a result, random forests typically achieve a high accuracy [9]. Moreover, random forests are very useful to handle datasets with many predictors [10]. Due to its efficiency in handling large datasets, we have chosen to apply Random Forest.

First, a Random Forest model and workflow was set up using the rand_forest() and workflow() functions and tuned mtry, trees, and min_n. Then, the recipe was added. Secondly, a grid was laid down. Since the dataset has 8 predictors, the range for mtry should not be smaller than 1 or larger than 8. Considering the scale of this dataset and the risk of RStudio running too long, levels were set to 2 and the maximum number of trees to 1000. We tuned the model and printed the results by using autoplot(). Thirdly, collect_metric() and arrange() were utilized to determine the AUC of the best-performing random forest model on the folds. Fourthly, finalize_workflow(), fit(), and augment() were leveraged to train and evaluate the best-performing random forest model on the training set and assess its performance on the testing set, thereby obtaining model accuracy. Lastly, the AUC value of the best-performing model was calculated via the testing set by using the augment( ) and roc_auc() functions. The accuracy of this model and its AUC value was recorded as study results.

2.8. *Boosted Tree*

Boosted tree represents a method in machine learning that achieves both regression and classification goals effectively [11]. It is an ensemble learning technique that combines multiple weaker prediction models (usually decision trees) to form a powerful ensemble model. In the past few years, this technology has become one of the most influential methods in data mining and predictive modeling [12]. It performs well in handling nonlinear relationships, making it suitable for handling the high-dimensional and complex medical datasets involved in this study.

First, set up a Boosted Tree by using the boost_tree() function, and tuned mtry, trees, and min_n. After setting up the Boosted Tree model, a workflow was created, and the recipe added. Secondly, the grid was laid down. Since there are 8 predictors, the range for mtry should not be smaller than 1 or larger than 8. Set levels = 2 was selected and the maximum number of trees was set to 2000 to prevent RStudio from running too long. Thirdly, the workflow was tuned with the tune_grid() function and an autoplot() of the results were printed. The ROC_AUC of the best-performing boosted tree model on the folds was identified using collect_metric() and arrange(). Fourthly, finalize_workflow(), fit(), and augment() were utilized to fit the best-performing boosted tree model to the training set, after which its performance on the testing set was assessed using augment() and roc_auc(). The accuracy of this model and its AUC value was recorded as study results.

2.9. K Nearest Neighbors

The K Nearest Neighbors (KNN) algorithm is a fundamental and widely used machine learning algorithm since it is easy to implement and has distinguished performance [13]. It is a supervised algorithm that predicts the classification of unlabeled data by considering the features and labels of the training data [14]. It works by measuring distance to find the K nearest neighbors of data points and making predictions based on the information of these neighbors. The algorithm is well-known its usage in solving both regression and classification challenges problems for different size, noise levels and label numbers [15]. K Nearest Neighbors was chosen for its efficiency in classification tasks.

A K Nearest Neighbors model was first set up using the nearest_neighbor() function, and the neighbors were tuned. Secondly, the workflow was set, and the recipe added. Secondly, a tuning grid was established. Thirdly, the model was tuned and an autoplot() of the results was printed. The collect_metric() and the arrange() functions were leveraged to identify the ROC_AUC of the best-performing KNN model on the folds. Fourthly, finalize_workflow(), fit(), and augment() were used to fit the best-performing KNN model to the training set, after which its performance on the testing set was evaluated using accuracy() and roc_auc(). The accuracy of this model and its AUC value was recorded as study results.

3. Results

After trying 6 different models, Logistic regression was deemed to be superior to LDA and QDA models. Table below exhibits the AUC and accuracy values of these 4 types of models on the testing set. The figure shows the ROC curves for these 4 types of models on the testing set.

Table 2. The AUC and accuracy values of Logistic Regression, Random Forest, Boosted Tree, and K Nearest Neighbors models on the testing set

Model Name	*Accuracy*	*AUC*
Logistic Regression	0.746	0.823
Random Forest	0.734	0.805
Boosted Tree	0.740	0.817
K Nearest Neighbors	0.720	0.788

Study results show that logistic regression is the best of the 4 models since it has the highest accuracy and AUC values.

4. Discussion

Although the logistic regression is the best model, the accuracy of the final model (logistic regression) is about 74.6%, which is less than ideal. There are three potential limitations that may have led to these results. First, the use of the undersamping technique during data processing. The original dataset is characterized by a significant imbalance, with 292,422 observations associated with non-heart disease patients, compared to only 27,373 observations related to heart disease patients. An undersampling technique was employed to randomly remove 265,035 observations without heart disease, thereby obtaining 27,387 observations of non-heart disease patients and 27,373 observations of heart disease patients. The deleted data constituted 90.6% of the original dataset’s observations without heart disease, and even if the undersampling technique’s removal of data was completely random, it could potentially result in the loss of critical information, ultimately leading to a reduction in the model’s accuracy.

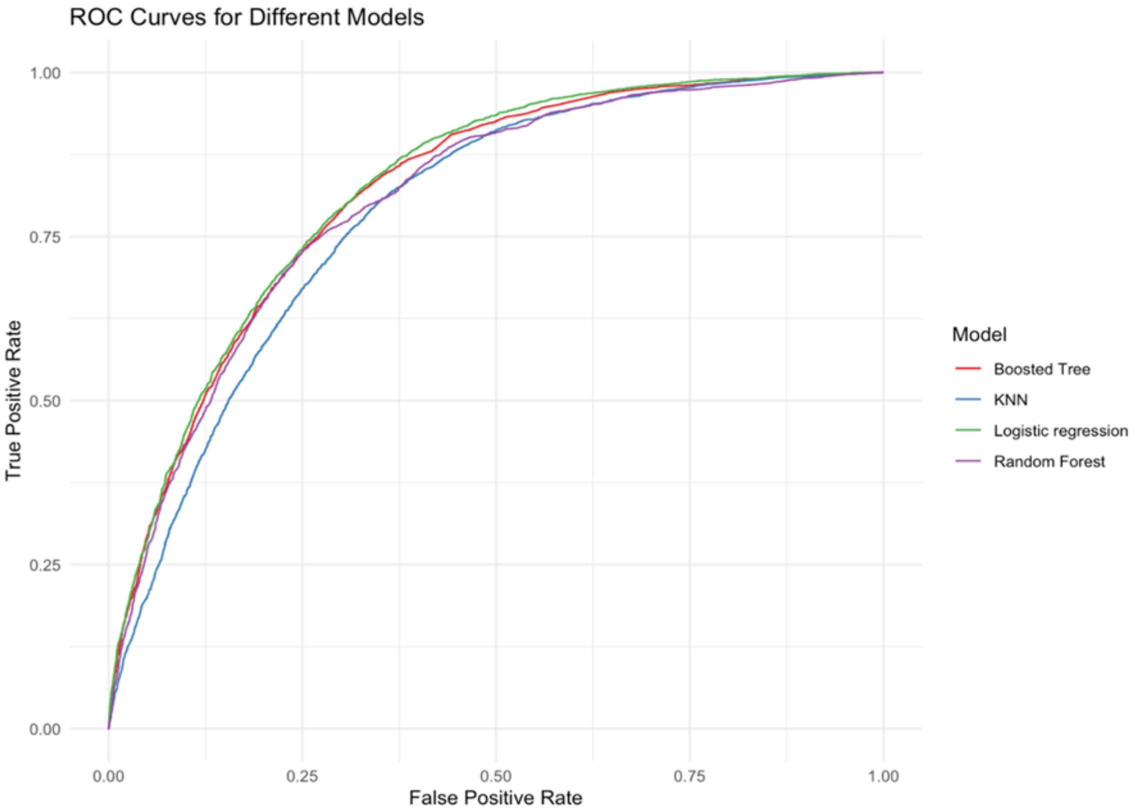

Figure 1. ROC Curves for different models

Secondly, limitations in feature selection. There were 17 risk factors in total in the original dataset, but only 8 was selected for this study. The deletion of the remaining 7 risk factors has also wiped out their relationship with patients' possibility of having heart disease, and that could result in reduced accuracy.

Thirdly, delineation of parameters during model development. To prevent RStudio from running for extended periods, the range of parameter trees for both Boosted Tree and Random Forest was restricted to only 10 to 1000 and 10 to 2000, respectively. This relatively limited value range may prevent Boosted Tree and Random Forest from achieving their maximum potential accuracy.

Therefore, future research may benefit from more sophisticated imbalance handling techniques, feature selection strategies, and model parameter adjustments, through which higher model accuracy may potentially be achieved.

5. Conclusion

This study demonstrates the potential of machine learning in predicting heart disease. In this project, a predictive model was generated to predict whether individuals will get heart disease based on key risk factors. Logistic regression emerged as the most accurate model among 6 standard machine learning techniques, including Logistic Regression, LDA, QDA, Boosted Tree, Random Forest, and K Nearest Neighbors. Although the final model achieved an accuracy rate of just 74.6%, it does highlight the importance of machine learning algorithms in predictive models for heart disease diagnosis and intervention. Further research is also necessary, though additional focus could be placed on addressing data imbalance, enhancing feature selection, and optimizing parameters. Should the accuracy of predictive models be improved, they will be well-poised to transform the field of public health. Countless lives would be saved.

References

[1] World Health Organization. Cardiovascular diseases (CVDs) [Internet]. World HealthOrganization. World Health Organization; 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics [Internet]. 2019 Mar [cited 2020 Feb 1]; 36:82–93. Available from: https://www.sciencedirect.com/science/article/pii/S0736585318308876

[3] G A, Ganesh B, Ganesh A, Srinivas C, Dhanraj, Mensinkal K. Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings [Internet]. 2022 Jun 1 [cited 2022 Jun 21]; 3(1):127–30. Available from: https://www.sciencedirect.com/science/article/pii/S2666285X22000449

[4] Pytlak K. Personal Key Indicators of Heart Disease [Internet]. www.kaggle.com. Available from: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[5] ABBOTT RD. LOGISTIC REGRESSION IN SURVIVAL ANALYSIS. American Journal of Epidemiology. 1985 Mar;121(3):465–71.

[6] Herman P, Prasad G, McGinnity TM, Coyle D. Comparative Analysis of Spectral Approaches to Feature Extraction for EEG-Based Motor Imagery Classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2008 Aug;16(4):317–26.

[7] Mahmodi K, Mostafaei M, Mirzaee-Ghaleh E. Detection and classification of diesel-biodiesel blends by LDA, QDA and SVM approaches using an electronic nose. Fuel. 2019 Dec;258:116114.

[8] Bhattacharyya S, Khasnobish A, Chatterjee S, Konar A, Tibarewala DN. Performance analysis of LDA, QDA and KNN algorithms in left-right limb movement classification from EEG data. 2010 International Conference on Systems in Medicine and Biology. 2010 Dec.

[9] Speiser JL, Durkalski V, Lee WM. Random forest classification of etiologies for an orphan disease. Statistics in Medicine. 2015 Feb 28;34(5):887–99.

[10] Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications. 2019 Nov;134:93–101.

[11] De’ath G. Boosted Trees for Ecological Modeling and Prediction. Ecology [Internet]. 2007 [cited 2023 Dec 20];88(1):243–51. Available from: https://www.jstor.org/stable/27651085

[12] Lee S, Kim JC, Jung HS, Lee MJ, Lee S. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomatics, Natural Hazards and Risk. 2017 Apr 10;8(2):1185–203.

[13] Zhang S, Li X, Zong M, Zhu X, Wang R. Efficient kNN Classification With Different Numbers of Nearest Neighbors. IEEE Transactions on Neural Networks and Learning Systems. 2018 May;29(5):1774–85.

[14] Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. Nature Methods. 2018 Jan;15(1):5–6.

[15] Zhang S, Li X, Zong M, Zhu X, Cheng D. Learning k for kNN Classification. ACM Transactions on Intelligent Systems and Technology. 2017 Apr 22;8(3):1–19.