Predictive Analysis of Heart Disease: Comparing Decision Tree, Random Forest, and XGBoost Models

Ruoxi Hao Yuxiao Nie Zhuoxin Fu

Introduction

According to the National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP), heart disease is one of the leading causes of death each year in the United States. Nearly half of Americans (47%) have at least one of the three major risk factors for heart disease: high blood pressure, high cholesterol, and smoking[1]. The Centers for Disease Control and Prevention (CDC) has stated other key indicators, including diabetes status, obesity (high BMI), not getting enough physical activity, or drinking too much alcohol. Besides these common factors, diseases such as angina[2], asthma[3], and kidney disease[4] may also have an impact on the development of heart disease.

Identifying individuals at increased risk of heart disease can significantly enhance prospects for delayed onset and improve follow-up treatments[5]. Therefore, predicting the potential risk of having heart disease based on known factors and distinguishing the prevalence of heart disease from symptoms of other common illnesses is crucial in healthcare. In turn, developments in computing allow the application of machine learning methods to detect "patterns" in the data that can predict a patient's condition[6].

This project aims to build prediction models to utilize indicators gathered from patient-level survey data to assess the likelihood of heart disease, offering diagnostic references and treatment basis for the prevention and treatment of heart disease.

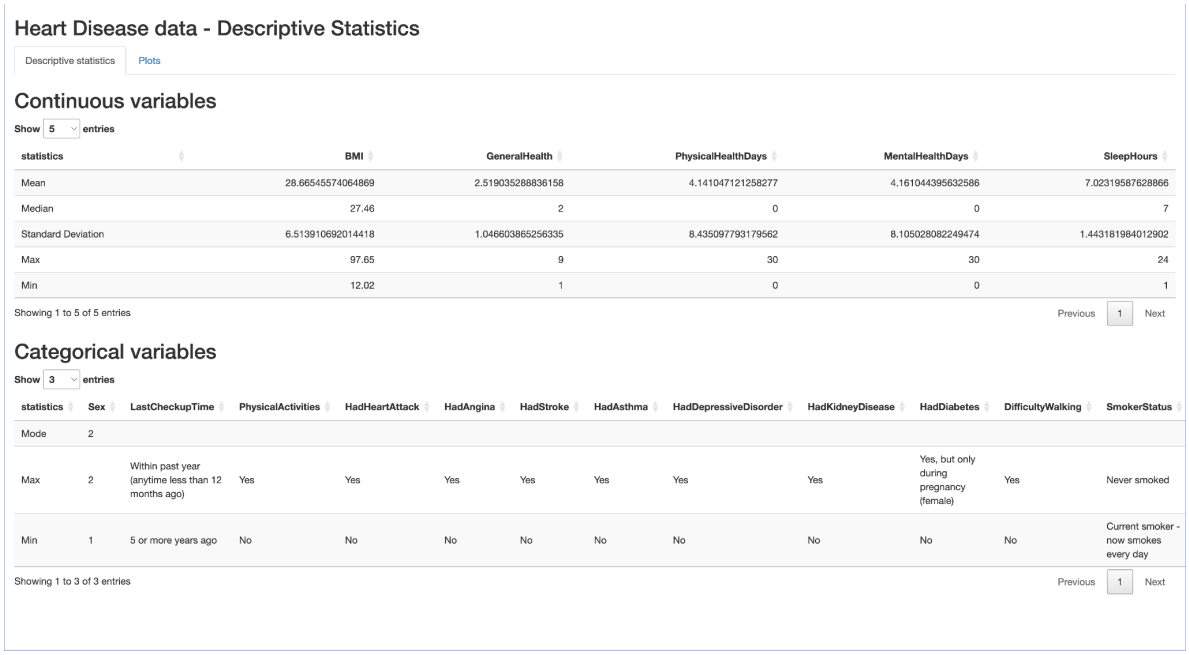
Data

Using the CDC data from the 2022 cycle of BRFSS (Behavioral Risk Factor Surveillance System) data collection with 445,132 participants and removing missing data for variables of interests, our study population included 249484 adult participants (≥ 18 years) from 50 states, the District of Columbia, Guam, Puerto Rico, and the US Virgin Islands.

According to the literature, the 25 variables we used to predict heart disease status are age, gender, bmi, race, state, sleep hours, self-reported general health condition, mental and physical activities, alcohol and smoking status, last checkup time, whether the person is disabled for walking, whether the person had been diagnosed with angina, stroke, asthma, depressive disorder, kidney disease, and diabetes.

We standardized data, handled outliers, cleaned unreasonable data, and mapped easy-to-read labels to improve data quality and interpretability. As the large volume of data may lead to computational complexities, we used H2O, a fast open-source machine learning and predictive analytics platform, to efficiently process and analyze our data set.

We built a Shiny app to describe the structure of the data set by showing descriptive analyses. The app comprises a sidebar operating as a control panel and a main display designated for the depiction of analysis tables and graphical parts. The descriptive statistics part facilitates an in-depth exploration of variables; this examination is divided based on the nature of the variables into continuous and categorical, as demonstrated in Figure 1. Within the graphical part, variables are also extracted based on their respective categories (continuous or categorical) and graphs of each variable, segregated into histograms, box plots, and barplots, are exhibited separately in Figure 2.



Heart Disease data - Descriptive Statistics

Descriptive statistics | Plots

Continuous variables

Show 5 entries

statistics	BMI	GeneralHealth	PhysicalHealthDays	MentalHealthDays	SleepHours
Mean	28.66545574064869	2.519035288836158	4.141047121258277	4.161044395632586	7.02319587628866
Median	27.46	2	0	0	7
Standard Deviation	6.513910692014418	1.046603865256335	8.435097793179562	8.105028082249474	1.443181984012902
Max	97.65	9	30	30	24
Min	12.02	1	0	0	1

Showing 1 to 5 of 5 entries — Previous 1 Next

Categorical variables

Show 3 entries

statistics	Sex	LastCheckupTime	PhysicalActivities	HadHeartAttack	HadAngina	HadStroke	HadAsthma	HadDepressiveDisorder	HadKidneyDisease	HadDiabetes	DifficultyWalking	SmokerStatus
Mode	2											
Max	2	Within past year (anytime less than 12 months ago)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes, but only during pregnancy (female)	Yes	Never smoked
Min	1	5 or more years ago	No	No	No	No	No	No	No	No	No	Current smoker - now smokes every day

Showing 1 to 3 of 3 entries — Previous 1 Next

Figure1: Descriptive statistics tables of variables

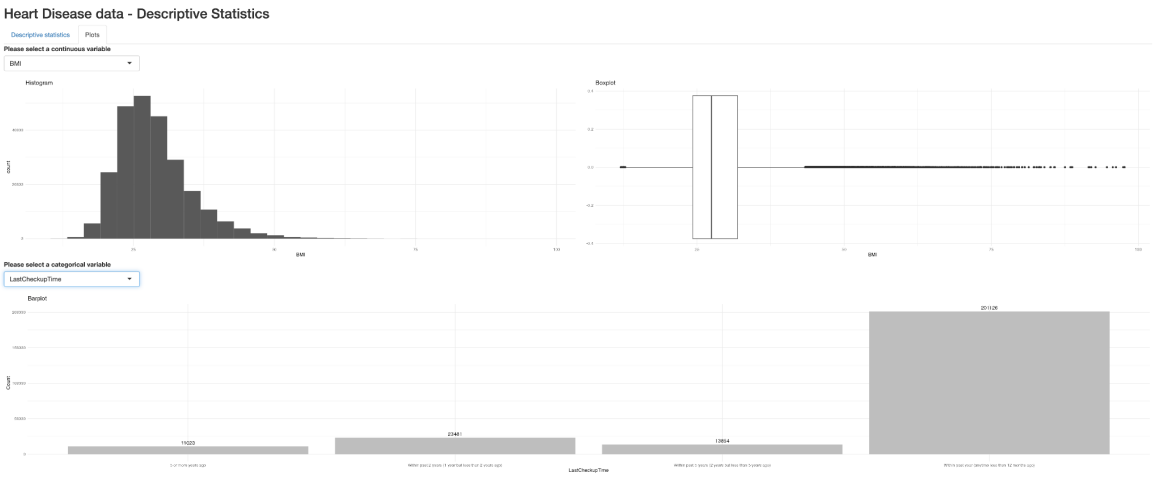


Figure 2: Graphs of variables

Models and results

Due to the huge amount of data we have, using the machine learning package that comes with R may cause insufficient memory, so we decided to use the h2o package, an in-memory, distributed, fast, and scalable machine learning platform, for modeling. We then chose to fit three different models, which are decision tree, random forest and XGBoost. We used 70% of the final data for training and 30% for testing.

We first fit a decision tree model. Decision trees are trees that contain nodes which split the data into a sample into homogeneous subsets, until a predefined stopping criterion is met. Decision tree machine learning algorithms decide the optimal way to split these nodes to best classify the data point. The overall AUC of this model is 0.8687, with specificity of 96.5% and sensitivity of 50.9%.

After fitting the decision tree model, we decided to fit a random forest model to further improve the accuracy of the prediction and reduce the overfitting of individual trees to create a more robust model. By fitting this ensemble model, it uses a collection of decision trees to classify data and then aggregates the votes from different trees to decide the final class of the test object. We expected to get better results in predictive performance. As a result, the AUC of the random forest model is 0.8777, indicating a good predictive performance. Analyzing the confusion matrix reveals that while our model has a high specificity of 96.5%, correctly identifying 96.5% of individuals without heart disease, it incorrectly identified 3.5% as having heart disease. On the other hand, the model exhibits a lower sensitivity of 49.8%, indicating that it correctly identified only 49.8% of individuals with heart disease and incorrectly classified 50.2% as not having heart disease.

Finally, we decided to use Extreme gradient boosting (Xgboost), a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library, to fit the model. It is an extension of gradient boosting, constructs a series of decision trees where each tree aims to rectify the errors of its predecessor. This iterative approach sets specific objectives for subsequent models, progressively refining predictions and ultimately leading to a robust and highly predictive model. We therefore fit a more sophisticated model using the Xgboost approach in order to obtain a more precise prediction result. The overall AUC of this model is 0.8889, suggesting that it's a more appropriate method to predicting the risk of having heart disease based on this dataset as expected. From the result of the confusion matrix, the specificity of this model is 97.9% and the sensitivity of this model is 52.4%. The high specificity indicates this model can correctly anticipate 97.9% of people who truly have heart disease and 52.4% people without heart disease. However, the relatively low sensitivity shows the model tends to misclassify 47.6% people with heart disease as those who do not.

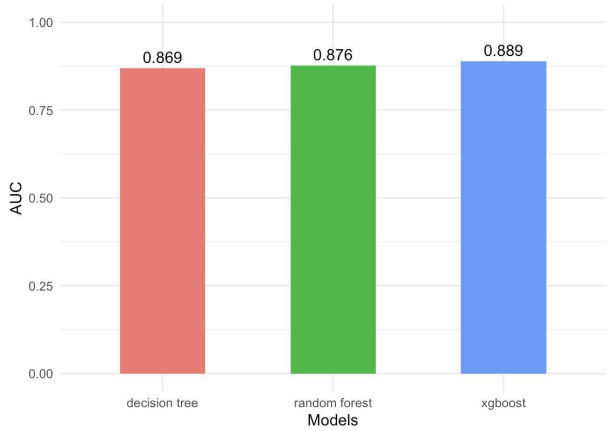


Figure 3: Accuracy of models

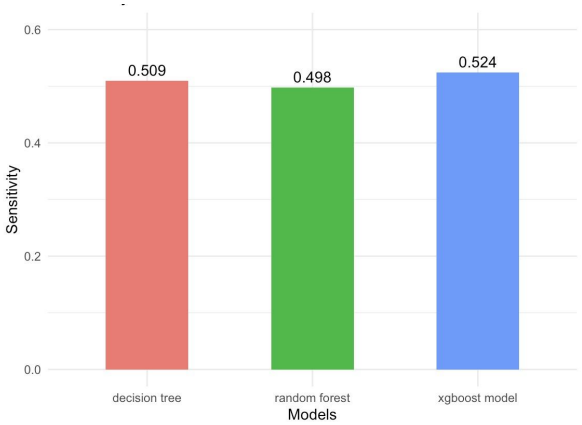


Figure 4: Sensitivity of models

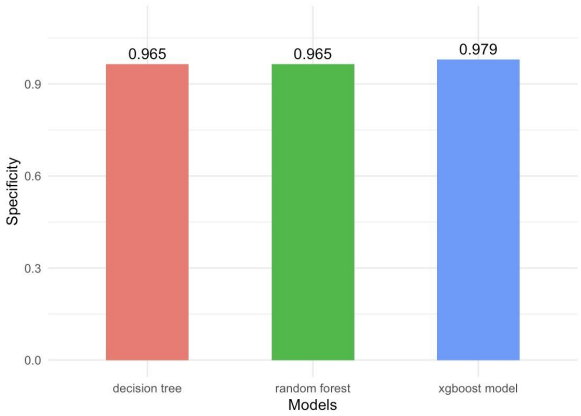


Figure 5: Specificity of models

Conclusion and discussion

The fitting results of the model reveal a progressive, albeit slight, increase in predictive accuracy across the three tested methods. Concurrently, all three models display a pattern of low sensitivity paired with high specificity in predicting heart disease. This trait implies a commendable performance in excluding heart disease cases, though subsidies considerably in accurately identifying real instances of heart disease. In essence, the models harbor a tendency toward false negatives, compromising detection rates, while subsequently exhibiting a diminished likelihood of false positive results. Consequently, for patients ascertained by the models as having the potential for heart disease, we strenuously advocate adjunctive screening tests to enhance diagnostic accuracy and ensure comprehensive patient care.

Besides, there are still some limitations for the models. When developing the model, an intrinsic challenge lies in the absence of temporal information pertaining to the onset or duration of heart disease in the training data. It is commonly recognized that patients' lifestyle modifications - induced by diagnosis - can substantially alter the disease's progression and the patient's health status. Such amendments, when not incorporated into the modeling process, can introduce systematic biases. In essence, the absence of this dynamic information and the potential changes in patients' behavior post-diagnosis likely constrain the model's capacity to accurately capture the complexity of heart disease progression and thereby may result in biased model predictions. In addition, there are some variables in models that may be reported by patients, so it is possible to have other systematic bias such as reporting bias or recall bias, leading to inaccuracy in predicting. Moreover, certain variables within the models might be derived from patient self-reporting, introducing the possibility for additional systematic biases, such as reporting bias or recall bias. These biases could compromise the veracity of the data, subsequently resulting in inaccuracies within the predictive capabilities of the models.

Reference

1, Fryar, C. D., Chen, T. C., & Li, X. (2012). Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010 (No. 103). US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

2, Liu, Y., Li, Z., Shen, D., Song, Y., Huang, M., Xue, X., Xie, J., Jiao, Z., Gao, S., Xu, Y., Gao, S., Wang, X., Xu, Q., Gao, S., Li, C., Li, L., Niu, K., & Yu, C. (2019). Adjuvant treatment of coronary heart disease angina pectoris with Chinese patent medicine: A prospective clinical cohort study. Medicine, 98(33), e16884.

3, Ingebrigtsen, T. S., Marott, J. L., Vestbo, J., Nordestgaard, B. G., & Lange, P. (2020). Coronary heart disease and heart failure in asthma, COPD and asthma-COPD overlap. BMJ open respiratory research, 7(1), e000470.

4, Sriperumbuduri, S., Clark, E., & Hiremath, S. (2019). New Insights Into Mechanisms of Acute Kidney Injury in Heart Disease. The Canadian journal of cardiology, 35(9), 1158–1169.

5, Wang, Z., Zhu, C., Nambi, V., Morrison, A. C., Folsom, A. R., Ballantyne, C. M., Boerwinkle, E., & Yu, B. (2019). Metabolomic Pattern Predicts Incident Coronary Heart Disease. Arteriosclerosis, thrombosis, and vascular biology, 39(7), 1475–1482.

6, Alanazi R. (2022). Identification and Prediction of Chronic Diseases Using Machine Learning Approach. Journal of healthcare engineering, 2022, 2826127.