Comparison of Models for Predicting Outcomes in Patients for Heart Disease

Hanzhang Liu*

Sendelta International Academy, Shenzhen, China

*Corresponding author: felix.hanzhang.liu@student.sendelta.com

Abstract. Over one quarter of all global deaths are caused by heart disease. Heart disease takes the lives of nineteen million people every year, making it the major cause of death. The risks of cardiovascular disease have been steadily rising across the globe over the past thirty years and COVID-19 has only exacerbated the predicament. Curing heart disease is a very expensive task, and thus a prediction is necessary. This study seeks to review the accuracy and accessibility of three models that are commonly used to predict decisions, which are Logistic Regression, Decision Tree, and Random Forest. Each method is used to fit models using data with 70000 cases and eleven variables. The accuracy rates for the Logistic Regression, Decision Tree, and Random Forest are 0.54, 0.56, and 0.51, respectively. Through a comparison, the Random Forest model is superior to the other models with the smallest misclassification error. Nevertheless, it does not have the accessibility that the other two regressions possess.

Keywords: Heart disease; Logistic regression; Decision tree; Random forest.

1. Introduction

Heart disease in its totality is the prominent cause of death worldwide. Cardiovascular disease (CVDs), the most prevalent type of heart disease, causes heart attacks and strokes. By itself, CVDs take 17.9 million lives a year worldwide [1]. Heart disease is an umbrella term that includes everything from inherited conditions to developed conditions. Altogether, heart disease is the cause for the deaths of 19 million people every year, causing 1 in 4 global deaths [2]. It ranks the third die prematurely for those under the age of 70 [1]. Behavioral risk factors of heart disease are tendencies such as nutritionless food consumption, lack of exercise, and tobacco or alcohol use. This can be seen in a rise of blood pressure and blood glucose, as well as overweight and obesity. These are all factors that the dataset will contain to model heart disease. These directly cause an increase in risk for heart attack, stroke, heart failure, and other heart complications. They can also be measured in primary care facilities and can be used to predict the likelihood of developing heart disease.

As societies develop, heart disease, especially cardiovascular disease, poses a large problem as obesity rates and mass consumption of unhealthy food rise [3]. In 2015, 19.5% of the adult population over some countries were obese, nearly one-fifth of all adult citizens [4]. This problem of obesity rates does not only plague developing countries, but also developing ones such as China and India. In China, the category of overweightness and obesity was present for more than more than 50% of adults and about 20% of children and adolescents, despite efforts in obesity prevention from the government [5]. The problem of obesity concerns all nations around the world and it contributes to the problem of cardiovascular disease. Adding on substance use in tobacco and alcohol [6], heart disease is a problem brewing within developed and developing countries.

Europe has stressed this disease as a primary concern within their society as it not only takes away many lives, but it also places a heavy burden on European economies that often provide free healthcare and other socialist policies [7]. In 2000, it was estimated that 74 billion euros are spent annually by European Union governments treating cardiovascular disease. Another 106 billion euros were lost from reduced productivity and the economic impact of illness and death. In the United States, heart disease and strokes cost America nearly 1 billion dollars a day in medical costs and lost productivity [8]. Treatment of cardiovascular disease raises a big economic toll for any government to take on. Moreover, these are the most developed countries with the most government income.

Developing countries that are also struggling with heart disease problems will have a hard time treating these diseases as they do not have the necessary economic resources to do so or are struggling with more serious health issues like malnutrition or communicable diseases [9].

The aim of this paper is to compare several different models for predicting outcomes in patients for heart disease. These papers have surprisingly high accuracy rates, much higher than the accuracy rates of the models that this paper provides. This is largely due to the small dataset that they were using, resulting in the risk of overfitting. The dataset that this paper uses is a larger dataset of 70,000 cases and 11 variables. This provides a more general look at the population of people and it reduces the likelihood of overfitting. The errors will be larger but it will be more generalized.

2. Methodology

This study aims to provide three comprehensive models on predicting the probability of heart disease. To achieve this, various machine learning algorithms were used, including Logistic Regression, Decision Tree, and Random Forest. The research data is examined through visualization to show how each factor impacts one another. Then each model will be put into training. Finally, an analysis of the accuracy of each of the models will be conducted.

In the past decade, global healthcare has started to utilize machine learning and big data. For those that concern cardiovascular disease, there have been multiple models which have a high accuracy. The model made by Narin et al. seeks to provide a model which can predict heart disease to improve the accuracy of the Framingham risk score [10]. It used data from 689 people who showed cardiovascular symptoms. It used a quantum neural network to build the model, resulting in an accuracy of 98.57%. Another study done by Alotaibi uses machine learning models to predict heart failure [11]. The paper aims to improve the accuracy of heart failure disease. The study used a dataset collected by the Cleveland Clinic Foundation. The decision tree method had the highest accuracy, 93.19%, surpassing previous prediction techniques of heart failure.

Machine learning has always played a role within the medical field. With machine learning models, predicting the possibility of heart disease would be possible. Those that have a high risk of heart disease can get warned beforehand and told to make lifestyle changes. This could prevent the disease from developing in the first place. Moreover, medical help and advice can also minimize this risk even more. This allows both developing and developed countries to solve the issue at hand without stressing their economies.

In this paper, three standard models commonly used to predict boolean outcomes are compared in their accuracy and qualitative counterparts. The model should allow medical professionals to better their therapeutic plans and explain needed lifestyle changes to those that require it depending on the possibility of risk. With a high accuracy in the model, it would provide hospitals and clinics with accurate information for treating individuals. Treatment of patients will be more streamlined and therefore more efficient. There are two specific questions this paper looks to address. First, how will the model predict heart disease? Second, what is the overall accuracy of the models? A comparison will be made through multiple errors as to which model would provide the best results.

2.1. Data

The dataset originated from Kaggle. This dataset includes a variety of factors and a large dataset. There have been other studies with extremely high accuracy rates but they often have small datasets from a localized area. This study attempts to provide a more general way to predict the possibility of heart disease by making it more accessible to predict. The dataset contains 70,000 patient records and 12 distinct features. The features do not require too much medical attention to be able to provide. The features include Age, Height, Weight, Gender, Systolic Blood Pressure, Diastolic Blood Pressure, Smoking, Alcohol Intake, Physical Activity, and Presence or Absence of Cardiovascular Disease. The variables, such as Gender and Alcohol Intake, are made into factors. The summary of the data is shown in Table 1.

Table 1. Three Scheme comparing

Feature	Variable	Type	Max, Min
Age	age	Integer	23713, 10798
Height	height	Integer	250, 55
Weight	weight	Integer	200, 10
Gender	gender	Factor	2, 1
Systolic Blood Pressure	ap_hi	Factor	-150, 16020
Diastolic Blood Pressure	ap_lo	Factor	-70, 11000
Cholesterol	cholesterol	Factor	3, 1
Glucose	gluc	Factor	3, 1
Smoking	smoke	Factor	1, 0
Alcohol Intake	alco	Factor	1, 0
Physical Activity	active	Factor	1, 0
Presence or Absence of Cardiovascular Disease	cardio	Factor	1, 0

In addition, Fig. 1 shows the essence of the data set and it show the correlation.

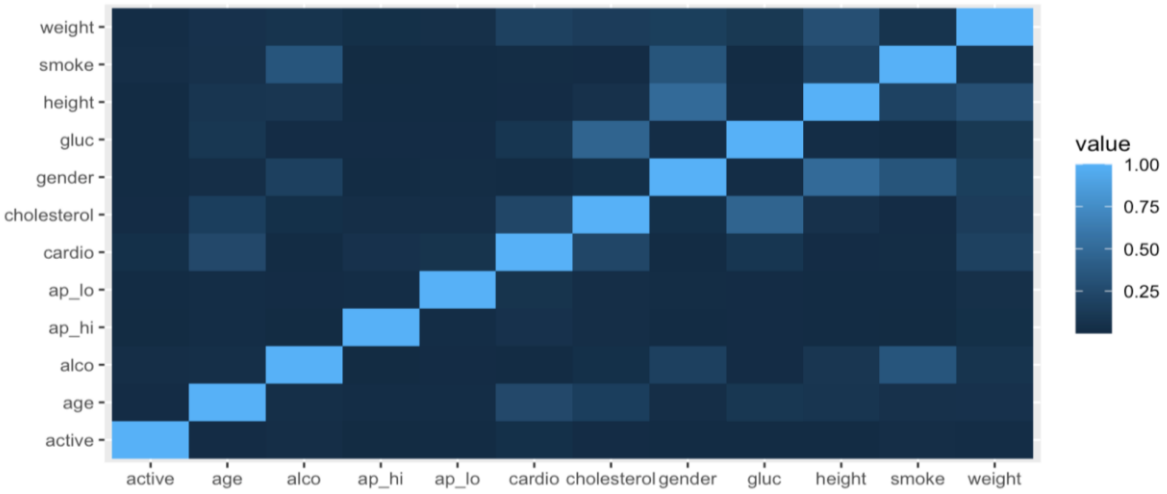

Fig. 1 Correlation map among different variables

2.2. Models

The Logistic Regression is used to find the probability of the success of an event in the shape of a logistic function. With any binary variable, Logistic Regression can be used. Despite the straightforward nature, it can be the best model depending on the dataset used. The disadvantage is that Logistic Regression assumes linearity between the dependent variable and the independent variable. This means that if two variables have little correlation or a non-monotonous relationship, Logistic Regression would have difficulty. Nonlinear problems cannot be solved using Logistic Regression. However, for the purposes of this study, Logistic Regression works.

The goal of a decision tree is to create rules that are inferred from the data structure. By creating branches of decisions, a decision tree classifies the data set by each decision. The algorithm is made to recursively separate data based on the most important feature of each node. Although not conventionally used for probability, it can be done by replacing the leaves with probability values. The biggest advantage for decision trees is that it is very easy for people to understand and use. However, decisions on probability are often poorly made compared to other models.

The Random Forest uses multiple decision trees and combines them into one result. The average of all the decision trees is taken as the prediction. It is easy to use for programmers as it does regression and classification simultaneously, and it is very likely to be more accurate than one

decision tree. The disadvantage is precisely its complexity as it is not very intuitive and would likely require some professional supervision to be able to use this model.

3. Results

After the training dataset and the testing dataset are created, the training dataset is used to train the model and the testing dataset is used to test for the accuracy of the model. With the extensivity of the dataset, the accuracy for the model should be relatively correct. The models of Logistic Regression, Decision Tree, and Random Forest will all be trained in this way. Next, a Misclassification Error accuracy test will be run on the models of the three models. With the code of the three models written, the results also come to light. A misclassification error is calculated with the testing data at the 1/3 prediction interval. These results are relatively straightforward and they answer the questions. Now, a comparison of these advantages and disadvantages will be discussed and the impact on heart disease will be examined.

3.1. Comparison of Accuracy

For the Logistic Function, the Misclassification Error is measured to be 0.54. Although this does seem a bit high, it must be noted that this is at the 1/3 prediction interval. For the Decision Tree, the Misclassification Error is measured to be 0.56. This is slightly worse than the Logistic Regression model in accuracy but in implementation the Decision Tree would likely be much clearer to an average person. Even at home, people without professional health knowledge could predict for themselves whether they might develop heart disease. A comparison is made to other models using the same dataset so the comparison should not be affected. For the Random Forest, the Misclassification Error is measured to be 0.51.

The error bounds are all pretty similar between the Logistic Regression (0.54), Decision Tree (0.56), and Random Forest (0.51) models. The Random Forest model has the highest accuracy but it is only marginally better than the other two models. The difference between the Random Forest model and the model with the lowest accuracy is only 0.05. The Logistic Regression model is second and it lies somewhat evenly in between the other two models being only 0.03 more erroneous than the model with the highest accuracy and 0.02 more accurate than the model with the least accuracy. The Decision Tree has the least accuracy.

3.2. Discussion

All three models have their own advantages and disadvantages. Due to the different natures of each model, they can be used in different settings so that each can cater towards their strong points in the prevention of heart disease.

The Random Forest model provides marginal benefits in its accuracy. Compared to the Decision Tree model, the Random Forest model should be able to provide much better accuracy yet is unable to show that. Moreover, a random forest is much less accessible in interpretation. When compared to the Logistic Regression model, the error difference contracts even more. The Logistic Regression is easy to use and can likely be able to run in any clinic or hospital of any size. A Random Forest requires a lot more calculation power that other models do not need, especially with a much larger dataset than the one that is used in this study. This would mean that perhaps it would require professional supervision or well-built medical facilities to have a well-trained Random Forest model. However, it should be noted that with a larger data set, the difference between the errors of the Random Forest model to the others will likely be larger than what is seen in this study as the Random Forest takes into consideration more things than the other models do.

The Logistic Regression model is proven to be almost as effective as the Random Forest model. The advantage of the Logistic Regression model is that it is easily accessible with any dataset. It is very efficient and does not require much time to train. Of course, it will likely be less accurate than the Random Forest model with bigger datasets but it is definitely a functional model for Heart Disease.

Perhaps on smaller domestic scales, a Logistic Regression would work better than the Random Forest model and the Decision Tree model.

The Decision Tree model is the least accurate of the three models. However, it is also the most accessible for the public. Decision Trees are very easy to understand and interpret because they have clear rules to determine the probability. Although it does have the lowest accuracy, it can be used to provide a generalized metric which the public can use with ease. With clear rules, the public can realize the risk of heart disease in their own homes. Many rural areas without good infrastructure lack hospitals. Even if there are hospitals, it is very far away from the surrounding villages. However, the world's poor have seen an increase in internet access. With internet access, the best Decision Tree could be spread as a metric so that the poor can check for themselves the risks for heart disease. It of course follows that this decision tree should be accurate and conveys the right information.

4. Conclusion

For developing nations who are struggling with more important health problems, an effective and accurate decision tree could be made and spread around the internet. Decision trees are easy to work with and suitable for public usage. The instructions are incredibly clear and easy to follow. This could serve both as a way to prevent heart disease and attract more attention to this problem. For local clinics or hospitals with localized data, it could be better to use a Logistic Regression model as it would likely apply most effectively to a small dataset. Data that is localized can be best reflected with a Logistic Function as larger more complicated machine learning models will simply result in overfitting and lower accuracy rates. The Random Forest model has the most potential to be an accurate model and should be utilized by researchers and medical professionals as a potential accurate predictor for heart disease. This holds the most promise as it can be easily used on a wider scale and still retaining the accuracy. It can help with generalized data on a large scale internationally. It is advised that health officials seek the Random Forest model as the most accurate model to predict heart disease within these three.

References

[1] Alotaibi Fahd Saleh. Implementation of Machine Learning Model to Predict Heart Failure Disease. International Journal of Advanced Computer Science and Applications, 2019, 10(6): 261-268.

[2] British Heart Foundation. Global Heart and Circulatory Diseases Factsheet, 2023.

[3] Bhurosy Trishnee and Jeewon Rajesh. Overweight and Obesity Epidemic in Developing Countries: A Problem with Diet, Physical Activity, or Socioeconomic Status?. The Scientific World Journal, 2014, 2014(1): 964236.

[4] OECD (2017), Behavioural Insights and Public Policy: Lessons from Around the World, OECD Publishing, Paris, 2017.

[5] Peng Wen, Zhang Jianduan, Zhou Haixia et al. The 2022 World Obesity Day and obesity prevention and control efforts in China. Global Health Journal, 2022, 6(3): 13-37.

[6] Hull Rodney, Mbele Mzwandile, Makhafola Tshepiso, et al. A multinational review: Oesophageal cancer in low to middle income countries (Review). Oncology Letters, 2020, 20(42): 45-56.

[7] Watson Rory. Heart disease rising in central and eastern Europe. BMJ: British Medical Journal, 2000, 320(3): 467.

[8] Stinson Claire. Heart Disease and Stroke Cost America Nearly $1 Billion a Day in Medical Costs, Lost Productivity. CDC Foundation, 2015.

[9] Thompson Karl. Why do developing countries have so many health problems?. Revise Sociology, 2021.

[10] Narin Ali, İşler Yalçın, and Özer Mahmut. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. Medical Technologies National Congress (TIPTEKNO), 2016, 2016(52): 1-4.