Heart Disease Diagnosis with Machine Learning

Yixuan Cheng, Jiamu Ding, Qian Yu

1 Introduction

1.1 Motivation and Application

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year (World Health Organization, 2021). The early detection is critical in the treatment of heart disease. We hope to build a binary classification model that can be useful in predicting the presence of heart disease based on the input parameters.

1.2 Prior Research

There is numerous research focusing on using machine learning models to do disease detection. Research by Fatima and Pasha (2017) provide the comparative analysis of different machine learning algorithms for diagnosis of different diseases. In another paper, the author utilized Support Vector Machine and other models to present an accurate model of predicting cardiovascular disease. (Dinesh, 2018). All of the papers give a guidance on different machine learning we can use.

1.3 Data Overview

The dataset used in this study is a collection of heart disease datasets from UCI Machine Learning Repository. It contains 918 patients' data with 11 features and a target variable, HeartDisease. Out of the total 918 patients in the data, 508 have been diagnosed with heart failure while 410 have not. Thus, the data are relatively balanced, which makes it useful for supervised learning.

2 Method

2.1 Exploratory data analysis

From histogram plots(Figure 1), we can observe the relationship of different categorical attributes related to heart disease. For gender types, males have a higher chance of having heart disease than females. For

exercise-induced angina, heart disease is often diagnosed when exercise-induced angina is present. For chest pain type, among patients with heart disease, the presence of asymptomatic chest pain is most common, followed by non-anginal pain and typical angina. For resting electrocardiogram results, results don't differ much. ST-T wave abnormalities can be a possible symptom of heart disease. For the slope of the peak exercise, people have a downward or flat ST Slope is often judged as having heart disease. And more

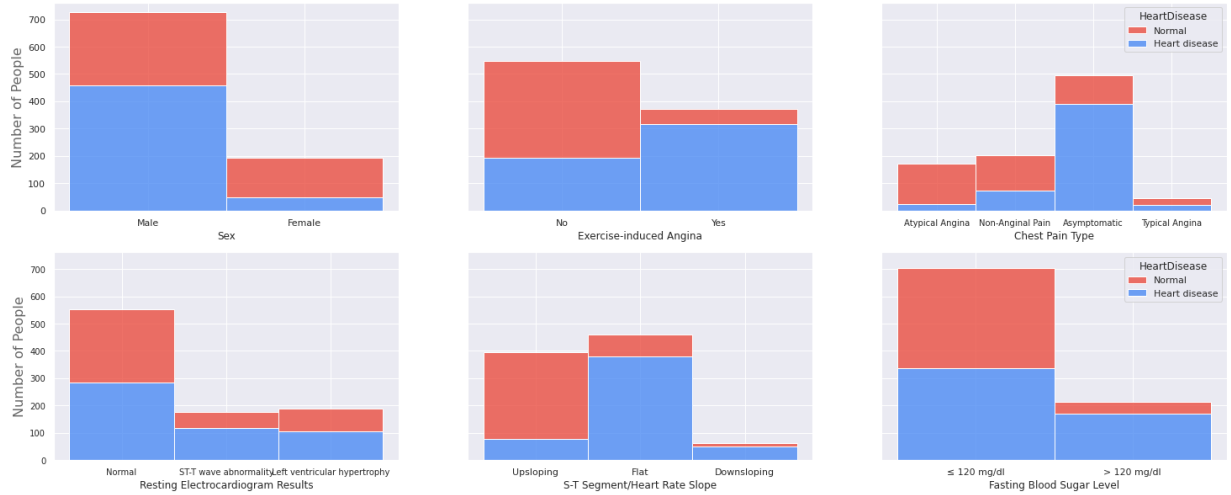


Figure 1: Categorical Features for Heart Disease

details about heartdisease with different parameters can be seen below(Figure 2).

• **Age:** The median age of patients with heart disease is 57, while people without heart diseases are younger with a median age of 51. There is a smaller spread in the boxplot for age of patients than normal people.

• **Systolic Blood Pressure:** The boxplots are proportional between the groups. The median blood pressure is about 130 mmHg in both groups.

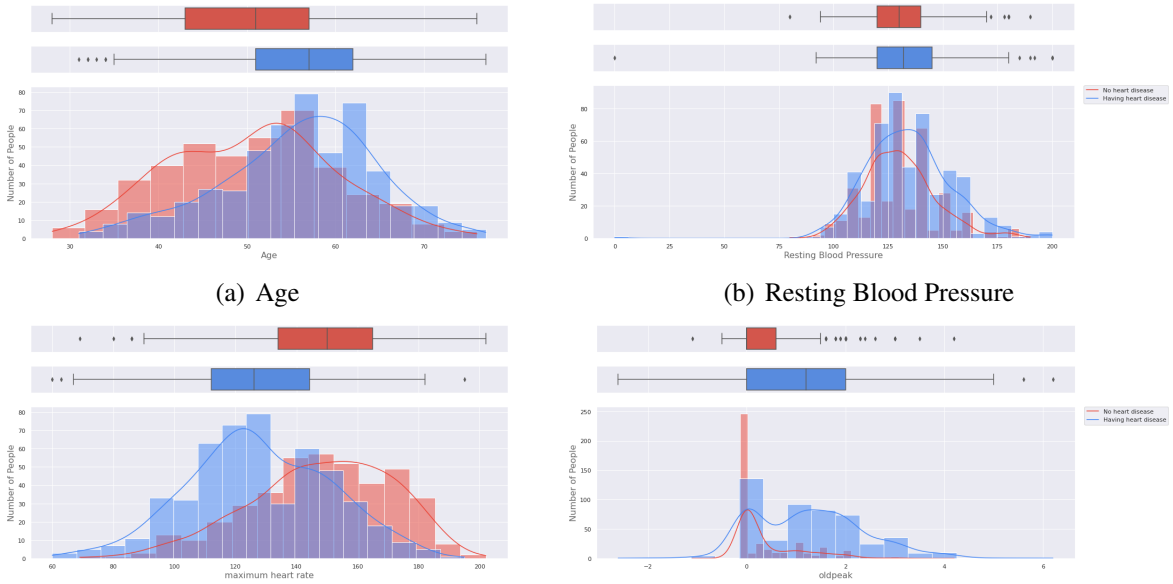
• **Heart Rate:** Patients without heart disease can reach higher maximum heart rates than patients with heart disease.

• **OldPeak:** There is a large difference in the distribution of ST segment depression between the groups. In patients with heart disease, the ST depression is more widely spread, and the patients also has higher oldpeak than normal people.

3 Processing Data and Model

3.1 Processing Data and Correlation Matrix

By analyzing the characteristics of the different features in the dataset, it can be seen that the parameters have different units and values, so they need to be normalized in the range from 0 to 1. The second step is to



(a) Age (b) Resting Blood Pressure

(c) Maximum Heart Rate (d) Oldpeak

Figure 2: Heart Disease with parameters

reduce the dimension of the categorical data. Since some categorical data have not only 0 and 1 values, but also possibly have multiple discrete values. So the next step is to disassemble the parameter with multiple possible values so that each variable is binary. These variables, in particular, are not easily involved in the calculation of the correlation coefficient, so they are not substituted in the process of correlation coefficient analysis. The calculation of the correlation coefficients is then performed to obtain Figure 3.

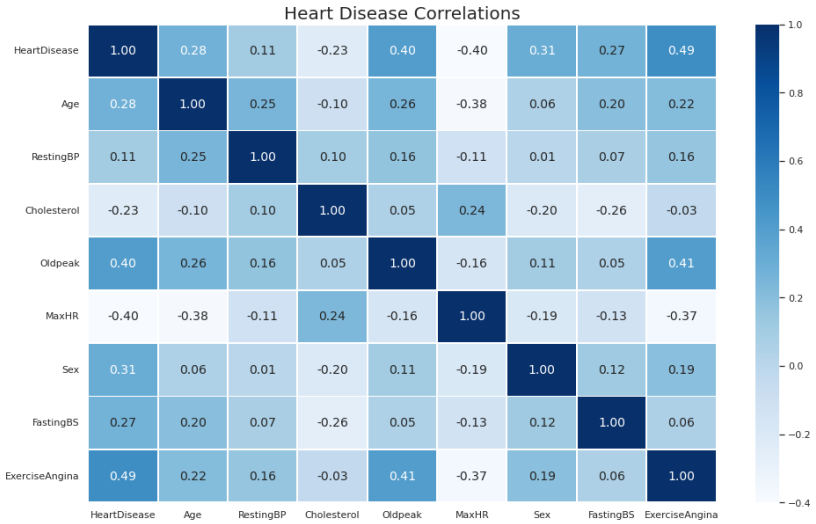


Figure 3: Heart Disease Correlation

Based on the correlations and scatterplots, HeartDisease has the strongest positive association to ExerciseAngina (correlation = 0.4) and the strongest negative association to MaxHR (correlation = -0.4). There

is also a moderately strong relationship between Oldpeak and Heartdisease of 0.40. As age increases, heart-disease rate tends to increase as well. There is also a strong relationship between these parameters that are directly associated with causing heartdisease.

3.2 Model

The next step is the selection and application of the model. In this study, three models were chosen for precision comparison in order to determine which model is more applicable in this case. First, the data were divided into training and validation sets in a ratio of 8:2. Then three classification algorithms, KNeighbors, RandomForest, and DecisionTrees, were selected for analysis and testing. The following results were finally obtained.

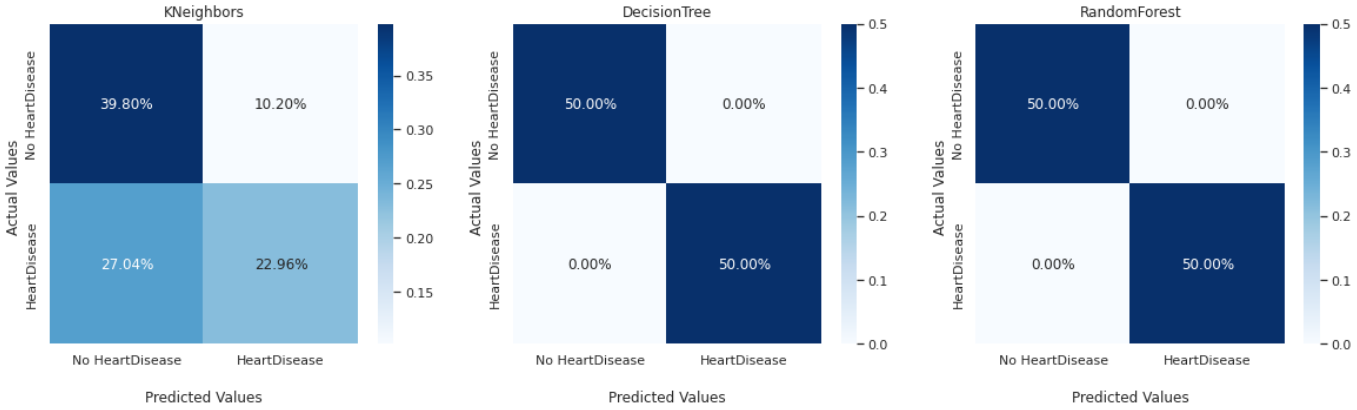


Figure 4: Three Models

As shown in the figure and table, the accuracy of KNeighbors' model is the lowest among the three algorithms, with an overall accuracy of 63% and an accuracy of only 69% and 60% for the cancer and non-cancer population, respectively. In contrast, RandomForest and DecisionTree both achieved 100 % accuracy.

4 Conclusion

Our models work really well after our preprocess. And the fully accurate results show that the output of RandomForest and DecisionTree algorithms are more reliable for the current data set which is not very large. At the same time, this is only for this case, in other cases, the background and conditions will change, so the operator needs to perform targeted testing.

References

1. Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., Singh, G. (2019, July). Prediction of coronary heart disease using machine learning: an experimental analysis. In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies (pp. 51-56).

2. Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-7).

3. Buczinski, S., Rezakhani, A., Boerboom, D. (2010). Heart disease in cattle: diagnosis, therapeutic approaches and prognosis. Veterinary journal (London, England : 1997), 184(3), 258–263.

4. Fatima, M., Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.