# Predicting Heart Disease

By: Xiuqi Li, Xiuqi Zheng, Zixin Yang

***Abstract* - This project is mainly focus on predicting people's heart disease. We use the dataset from UCI website, which include 13 features such as age, sex, maximum heart rate, to predict whether people will suffer heart disease. We use five methods to model the data and analyze the results: Support Vector Machine model (SVM), Random Forest Classifier model, Stochastic Gradient Descent model (SGD), Nearest Centroid Classifier (NCC) and K Nearest Neighbor Classifier model (KNN), and Logistic Regression model. By changing the proportion of training and testing dataset and the combination of feature inputs, we get each model's best performance. The higher the score, the better the model. According to the score of each model, we think KNN is the best model to predict the heart disease.**

## I. INTRODUCTION

Heart disease is a kind of chronic disease that threatens people's health in the world. It is the leading cause of deaths for men, women, and people of most racial and ethnic groups in the United States. One person dies every 36 seconds in the US from cardiovascular disease.[1] It also causes 4 out of every 10 deaths in the United States. This is more than all kinds of cancer put together. [2]

In all, there are two types of heart disease. One is people born with heart disease, called congenital heart disease. If people get heart disease later, called acquired heart disease, which is the most common kind of heart disease. The three most common kinds of acquired heart disease are coronary artery disease, congestive heart failure, and the bad heart rhythms. The coronary artery disease is a problem with the blood vessels that deliver blood to the heart muscle. The thin and weak blood vessel will cause sick and weak heart muscle. Therefore, causes the dysfunction of heart. The congestive heart failure is a condition that the heart is not pumping at normal levels. It may cause by the abnormal heart valves and weak heart muscle. The bad heart rhythm is a problem with electrical activity in the heart. The abnormal rhythm can cause abnormal blood moving and contractions in heart. [2]

The causes of heart disease can be varied. Heart disease occurs when plaque develops in the arteries and blood vessels that lead to the heart. This blocks important nutrients and oxygen from reaching your heart. Plaque is a waxy substance made up of cholesterol, fatty molecules, and minerals. Plaque accumulates over time when the inner lining of an artery is damaged by high blood pressure, cigarette smoking, or elevated cholesterol or triglycerides. Several risk factors play the important role in determining the possibility to develop heart disease. Two of these factors are age and heredity. The risk of heart disease increases around the age of 55 in women and 45 in men. People

have close family members who have a history of heart disease have higher rate of getting heart disease. Other risk factors for heart disease include obesity, insulin resistance or diabetes, high cholesterol and blood pressure, family history of heart disease, being physically inactive, smoking, eating an unhealthy diet, clinical depression. [3]

## II. TASK DESCRIPTION

The object of this project is to predict whether a patient is likely to get a heart disease or not. And the problem is a supervised machine learning linear problem, the target is noted by classification (1 if the patient has a heart disease, or 0, if not).

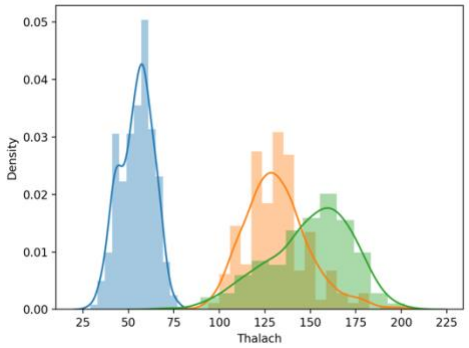
The data set of heart disease condition comes from the UCI website. Two methods (Logistic Regression, Support Vector Machine) are assigned to construct the model. The best model which has the highest score will be used to predict whether a patient is likely to get a heart disease or not.

*Dataset description*

The initial database contains 303 valid samples. Each sample has 13 features: age, sex (1=male, 0 = female), cp (chest pain type), trestbps (resting blood pressure), chol (serum cholestoral in mg/dl), fbs (fasting blood sugar > 120 mg/dl), restecg ( resting electrocardiographic results), thalach ( maximum heart rate achieved), exang(exercise induced angina, 1=yes, 0 = no), oldpeak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment), ca (number of major vessels (0-3) colored by flourosopy), thal (3=nomal, 6=fixed, 7=reversable defect).

The output is a label which shows if the people have a heart disease or not (1=yes, 0 = no).

*Virtualization of data*




Among these features, we choose three of them which we think that will contribute most to the prediction of heart disease: age, trestbps, and thalach. The blue part represents age, the orange part represents trestbps, and the green part represents thalach.

## III. MODEL & ANALYSIS

***Support Vector Machine***

The first method we use is a support vector machine (SVM). An SVM is a discriminative classifier defined by a separating hyperplane (Support Vector Machines, n.d.). Given labeled data, the SVM algorithm will output an optimal hyperplane which is used to categorize new samples. If in a two-dimensional space the hyperplane is a line dividing the feature space in two parts. In our case, the feature space is three-dimensional and SVM finds the optimal hyperplane to separate out the classes. We use the SVM function in the sklearn library.

*Benchmark*

First, we split the data into 152 training samples and 151 testing samples. We use parameters linear

kernel and C equal to 1000. C tells the SVM optimization method how much we want to avoid misclassifying each training sample. For larger values of C, the optimization chooses a smaller margin hyperplane if that hyperplane does a better job of getting all the training observations classified correctly (Support Vector Machines, n.d.). On the other hand, a small value of C will cause the optimization to look for a larger margin separating hyperplane, even if that hyperplane misclassifies more of the training observations (Support Vector Machines, n.d.). Using this method and parameters we get an accuracy score of 0.6639 and the following measures:

| **Confusion Matrix and Accuracy Measures** | | | | | |
|---|---|---|---|---|---|
| | | Predicted | | Precision0 | 0.72 |
| **Actual** | | y=0 | y=1 | Precision1 | 0.63 |
| | x=0 | 29 | 30 | Recall0 | 0.49 |
| | x=1 | 11 | 52 | Recall1 | 0.83 |

We can see that the precision of status 0 is 0.72. This indicates that of the total number of times our model predicted status 0, it was correct 72% of the time. And our model correctly predicted status 1 with probability of 63%. This model seems like predicting status 0 and status 1 very well. The recall of status 0 is 0.49, which is not very well. And we could see that the recall of status 1 is 0.83, which is comparatively well.

***Random Forest***

Random Forest belongs to the Bagging Algorithm. When a new sample is put into the model, everything decision tree makes decision and classification based on the sample's features. This method is based on the Random Forest Classifier module from scikit-learn. The key parameter of this method is "n_estimators". It represents the maximum amount of the weak learners (the amount of the decision trees). If this value is too small, the model is underfitting. If the value is too high, the calculation amount will be large and will no longer has significant increase. Overfitting may also occur if this value is too high.

*Benchmark*

Using this method, we set 'n_estimators' as 20 and get an accuracy score of 0.6475 and the following measures:

| **Confusion Matrix and Accuracy Measures** | | | | | |
|---|---|---|---|---|---|
| | | Predicted | | Precision0 | 0.68 |
| **Actual** | | y=0 | y=1 | Precision1 | 0.63 |
| | x=0 | 30 | 29 | Recall0 | 0.51 |
| | x=1 | 14 | 49 | Recall1 | 0.78 |

We can see that the precision of status 0 is 0.68. This indicates that of the total number of times our model predicted status 0, it was correct 68% of the time. And our model correctly predicted status 1 with probability of 63%. This model seems like predicting status 0 and status 1 comparatively well. The recall of status 0 is 0.51, which is not very well. And we could see that the recall of status 1 is 0.78, which is comparatively well.

And to make a comparison between SVM and Random Forest method, we compare the accuracy score and we get

the results that SVM Score is 0.6639 and RF Score is 0.6475. Therefore, here SVM has better performance.

***Stochastic Gradient Descent***

Stochastic Gradient Descent (SGD) method is a very useful and popular way to optimize the loss function and the neural network in the machine learning. This method reduces many computational burden and complex process, which makes the whole process more efficient and easier to implement, although it is more sensitive to the feature scaling.

*Benchmark*

To use scikit learn logistic regression method to train the data, we split the dataset into two equal size group for training and testing first, which means training set will contain 151 sample and testing set will contain 152 sample. Choose loss function by set "loss" parameter as logistic regression. Using this method and parameters we get an accuracy score of 0.5573 and the following measures:

| Confusion Matrix and Accuracy Measures | | | | | |
|---|---|---|---|---|---|
| | | **Predicted** | | Precision0 | 1.00 |
| **Actual** | | y=0 | y=1 | Precision1 | 0.54 |
| | x=0 | 5 | 54 | Recall0 | 0.08 |
| | x=1 | 0 | 63 | Recall1 | 1.00 |

We can see that the precision of status 0 is 1.00. This indicates that of the total number of times our model predicted status 0, it was correct 100% of the time. And our model correctly predicted status 1 with probability of 54%. This model seems like predicting status 0 and status 1 comparatively well. The recall of status 0 is 0.08, which is not very well. And we could see that the recall of status 1 is 1.00, which is very great performance.

***Nearest Centroid Classifier***

The nearest centroid classifier (NCC) gives the best K nearest neighbor model (KNN) score, which is a very simple and popular algorithm, because it can solve both regression and classification problem. It is easy to implement and simple to get the result. The principle and algorithm behind the methods is to find K number of training samples closest in distance to the new testing point, and predict the label from these by majority voting. The distance can, in general, be any metric measure: standard Euclidean distance is the most common choice.

*Benchmark*

Each time we use KNN model, we want to get the highest score and find the K value which could give us the highest score, so that at each iteration we run a function to find the highest score of KNN. Using this method and parameters we get an accuracy score of 0.6803 and the following measures:

| Confusion Matrix and Accuracy Measures | | | | | |
|---|---|---|---|---|---|
| | | **Predicted** | | Precision0 | 0.67 |
| **Actual** | | y=0 | y=1 | Precision1 | 0.69 |
| | x=0 | 39 | 20 | Recall0 | 0.66 |
| | x=1 | 19 | 44 | Recall1 | 0.70 |

We could see that the precision of status 0 is 0.67. This indicates that of the total number of times our model predicted status 0, it was correct 67% of the time. And our model correctly

predicted status 1 with probability of 69%. This model seems like predicting status 0 and status 1 comparatively well. The recall of status 0 is 0.67, which is also perform well. And we could see that the recall of status 1 is 0.69 which is also a comparatively good performance.

*Change the number of features*

To clarify how many features or inputs we use that could get the highest score, we apply many times of the model. In our dataset, we totally use three features to predict the label. We try different combinations and the summary table is shown below.

| Combination of features | KNN score |
|---|---|
| **Age** | 0.64473684 |
| **Trestbps** | 0.61184210 |
| **Thalach** | 0.67105263 |
| **Age+ Trestbps** | 0.67105263 |
| **Age+ Thalach** | 0.70394736 |
| **Thalach + Trestbps** | 0.69078947 |
| **Age + Thalach + Trestbps** | 0.69078947 |

From the result of the table, we find that the combination of age and thalach give us the highest score and trestbps gives us the lowest score. This indicates that when we predict the KNN result, thalach (the maximum heart rate achieved) is the most important feature that we should focus on, and only using trestbps cannot get an accurate prediction.

*Change the training sample size*

Then, we want to make clear that if the size of the sample will affect the result of prediction, so we change the sample size to see which size can give the best result. We change the proportion of the size of training and testing sample. The summary table is shown below.

| Proportion of testing sample | Accuracy (KNN score) |
|---|---|
| **0.5** | 0.69078947 |
| **0.4** | 0.68032786 |
| **0.3** | 0.71428571 |
| **0.2** | 0.75409836 |
| **0.1** | 0.87096774 |

According to the result in the table, we find when proportion is 0.1, we have the highest KNN score and when the proportion is 0.4, we have the lowest score. And the model roughly has the trend, that the fewer the data that testing sample has or the more the data that training sample has, the better the score.

***Logistic Regression***

Logistic regression is a function in the scikit-learn library. It is a classification algorithm in machine learning. It is very useful because this method could predict the probability of a classification variable, which means it could formulate the classification problem as a regression one.

*Benchmark*

To use scikit learn logistic regression method to train the data, we split the dataset into two equal size group for training and testing first, which means training set will contain 151 sample and testing set will contain 152 sample. After directly applying the logistic regression function of sklearn library, then we got the score for scikit learn. As a benchmark score, the scikit learn score is 0.63.

*Change the number of features*

To clarify how many features or input we use that could get the highest score, we apply the model many times. In our dataset, we totally use three features to predict the label. We try different combinations and the summary table is shown below.

| Combination of features | Scikit learn score |
|---|---|
| **Age** | 0.58552631578947 |
| **Trestbps** | 0.57236842105263 |
| **Thalach** | 0.66447368421052 |
| **Age+ Trestbps** | 0.56578947368421 |
| **Age+ Thalach** | 0.66447368421052 |
| **Thalach + Trestbps** | 0.65789473684210 |
| **Age + Thalach + Trestbps** | 0.63157894736842 |

From the result of the table, we find that the combination of age and thalach and only including thalach give us the highest score and the combination of age and trestbps gives us the lowest score. This indicates that when we predict the scikit learn result, thalach which means the maximum heart rate achieved is the most important feature that we should focus on, and only using age and trestbps cannot get an accurate prediction.

*Change the training sample size*

Then, we want to make clear that if the size of the sample will affect the result of prediction, so we change the sample size to see which size can give the best result. We change the proportion of the size of training and testing sample. The summary table is shown below.

| Proportion of testing sample | Accuracy (Scikit-learn score) |
|---|---|
| **0.5** | 0.631578947368421 |
| **0.4** | 0.62295081967213 |
| **0.3** | 0.63736263736263 |
| **0.2** | 0.63934426229508 |
| **0.1** | 0.77419354838709 |

According to the result in the table, we find when proportion is 0.1, we have the highest scikit-learn score and when the proportion is 0.4, we have the lowest score. And the model roughly has the trend, that the fewer the data that testing sample has or the more the data that training sample has, the better the score

***Model Comparison***

In this project, we totally implemented five forecasting Methods: Support Vector Machine, Random Forest, Stochastic Gradient Descent, K-Nearest Neighbors and Nearest Centroid Classifier, and logistic regression. For benchmark model, support vector machine method gives us an accuracy rate of 66.39%. Stochastic Gradient Descent method have an accuracy rate of 55.73%. K-Nearest Neighbors and

Nearest Centroid Classifier have an accuracy of 69.9% and 68.03%, respectively. Logistic Regression have an accuracy of 63.16%.

So, we can conclude that NCC method provides us the best prediction performance. And we could have a basic result of these five methods:

| LR | SVM | RF | SGD | NCC |
|---|---|---|---|---|
| 0.6316 | 0.6639 | 0.6475 | 0.5573 | 0.6803 |

## IV. CONDLUSION

Heart disease is a very serious disease that more and more people are facing. It is the leading cause of deaths for men, women, and people of most racial and ethnic groups in the United States. According to the survey, 40% of the deaths are due to heart disease. Whether it is congenital disease or acquired heart disease, there are obvious manifestations of peripheral sudden changes in blood pressure and heart rate, so in this project, we examine 303 cases of patients who have undergone heart disease. We use the following three features to predict patients' survival status: age of patient at time of operation, maximum heart rate achieved, and resting blood pressure.

The SVM method score is 0.6639. The random forest classifier method score is 0.6475. The SGD method score is 0.5573. Using KNN method and parameters we get an accuracy score of 0.6803. The Logistic Regression score is 0.63. And after model comparison, we find that KNN model using all three features and a training size of 50% and a testing size of 50% is the model that provides the highest accuracy, which is 69.9%.

We could know that the accuracy for these models are not high enough, the highest is still under than 70%. We could find the reason from the chosen variables. In actual world, we will have more reasons than these three parameters, which also could influence patients' heart disease conditions. Therefore, in the future, we could try to find more features and add them into our prediction.

## Reference


[1] Heart Disease Facts. (September 8, 2020). Centers for Disease Control and Prevention. Retrieve from https://www.cdc.gov/heartdisease/facts.htm

[2] Heart Disease. (n.d.). Wikipedia. https://simple.wikipedia.org/wiki/Heart_disease

[3] Causes and Risks of Heart Disease. (n.d.). Healthline. Retrieve from https://www.healthline.com/health/heart-disease/causes-risks#risk-factors