IST 707 Group Project

Hongdi Li & Yiqin Xu

Dec 9th, 2022

The Prediction of Heart Disease

Background

According to the CDC, heart disease has become the number one killer of human health. There are 17 million people who die of heart disease in the world every year, and one out of every three people who die from the disease is because of heart disease, so heart disease has also become the leading cause of human death. It is also a cause of death for most ethical, including African Americans, American Indians, Alaska Natives, Hispanics, and whites. About half of Americans (47%) have at least one of the 3 major risk factors for heart disease: high blood pressure, high cholesterol, and smoking. And there are also some invisible hazards in our life that can also become one factor leading to heart disease. Other key indicators include diabetes status, obesity (high BMI), lack of adequate physical activity, or excessive alcohol consumption [1]. Unlike the flu, heart disease doesn't seem to be detected in time, it is a chronic disease. For example, flu patients are often accompanied by fever, cough and other easily detectable physical body uncomfortable. However, sometimes heart disease may be “silent” and not diagnosed until a person experiences signs or symptoms of a heart attack, heart failure, or arrhythmia.[2] Thus, the purpose of this project is to use machine learning to predict whether people are likely to get sick and study the important factors that lead to heart disease.

Introduction

The collated dataset we selected comes from Kaggle, it’s called the 2020 annual CDC survey data of 400k adults related to their health status. Originally, the dataset came from the Centers for Disease Control and Prevention as part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts an annual telephone survey to collect data on the health status of U.S. residents. As the CDC describes: "BRFSS was established in 15 states in 1984 and now collects data from all 50 states plus the District of Columbia and three U.S. territories. [1]. It consists of 401,958 rows and 279 columns. A Kaggle user KAMIL PYTLAK removed some unimportant parameters from the research data . Reduce the data from nearly 300 columns to 18 columns, including BMI, age, smoking, drinking, etc. We will use this data collated by KAMIL PYTLAK for our research. The purpose of this project is to find out which of these variables have a significant impact on the likelihood of having heart disease.

Preprocessing

Considering our data size requirements (less than 100 example, we decided to randomly select 100 rows in the KAMIL PYTLAK dataset, and ensure that the people in the 50 rows of data have heart disease, and the people in the 50 rows do not suffer from heart disease. At the same time, keep all 18 columns of parameters, including:

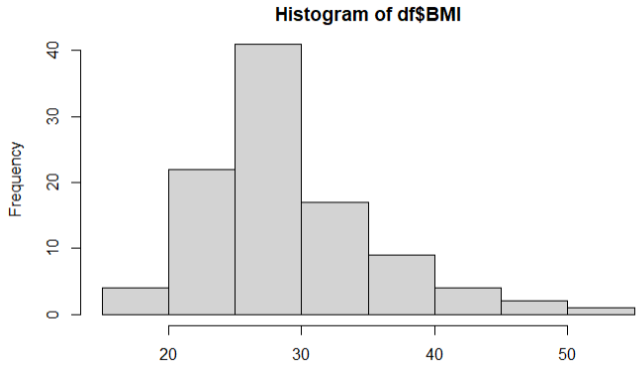
- **BMI** (body fat),

- **Smoking** (whether smoking),
- **AlcoholDrinking** (whether drinking alcohol),
- **Stroke** (whether stroke),
- **PhysicalHealth** (how many days people have physical health problems in the past 30 days),
- **MentalHealth** (how many days people have mental health problems in the past 30 days),
- **DiffWalking** (whether people have difficult in walk),
- **Sex**,
- **AgeCategory**,
- **Race**,
- **Diabetic** (whether people have diabetes),
- **PhysicalActivity** (whether people have had physical activity in the past 30 days),
- **GenHealth** (evaluation of self-health),
- **SleepTime** (length of sleep),
- **Asthma** (whether people have asthma),
- **KidneyDisease** (whether people have kidney disease),
- **SkinCancer** (whether people change skin cancer).

After discussion, we found that we still need to classify some variables in the data set, because the values of some parameters in the data set are too specific. Considering that our data set only has 100 rows, too specific values may not be good for the research effect. Therefore, we classify and reassign some parameters in the data set, including BMI, SleepTime, PhysicalHealth, and MentalHealth. And name the values in other parameters, for example, replace "No" in the Smoking column with "not smoke". The real-world problem of this project could be what are the different factors/problems that directly or indirectly affect heart disease? We are mainly focus on the prediction of the factors that cause heart disease and the prediction when people will get sick.

Descriptive Analysis

Considering that we want to classify the values in some columns, we need to perform analysis on these columns. Since this data set is too large, we randomly sampled 100 rows of data for analysis.



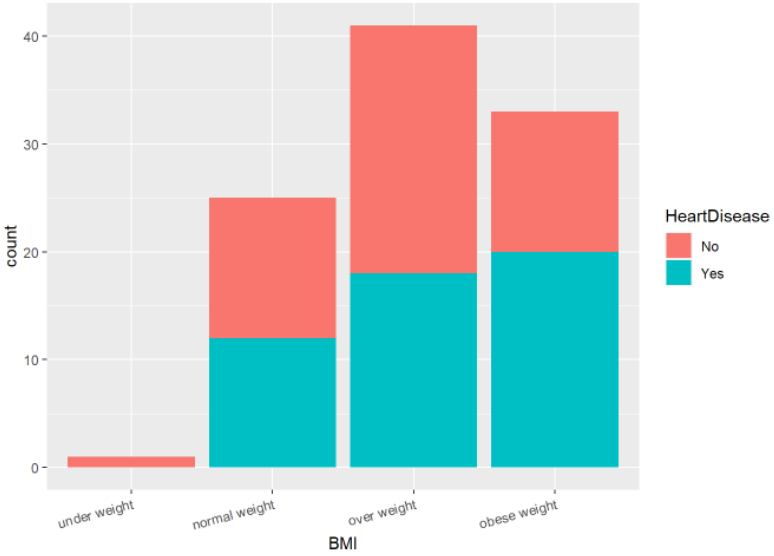
BMI

In BMI column, the lowest value is 19.97, the highest value is 52.09, and the average value is 29.02. By constructing a histogram image, the image has a positive skewed distribution.

(Histogram of BMI)

According to CDC's description of BMI, Below 18.5 is Underweight, 18.5 – 24.9 is Healthy Weight, 25.0 – 29.9 is Overweight, 30.0 and Above is Obesity [3]. We combined CDC's standards with our data information, we classify the number in BMI column into 4 indicators, reassign them in "underweight", "normal weight", "over weight" and "obese weight". According to statistics, among the 100 subjects we studied, 1 was underweight, 25 in normal BMI, 41 were overweight, and 33 were obese.

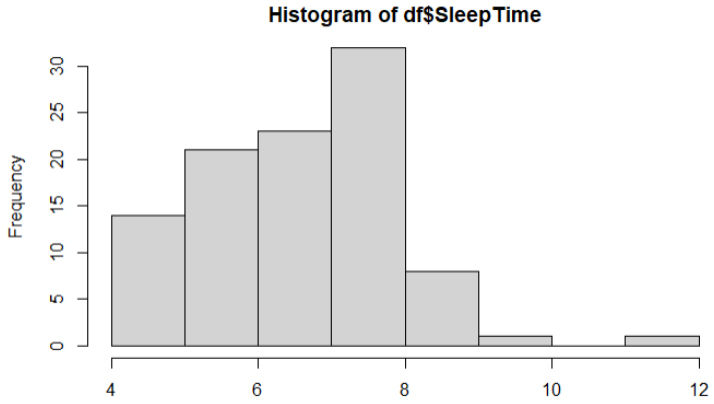
(Histogram of BMI with after classified value)



Re-analyzing the BMI column, we found that as "BMI went up," there seemed to be a greater risk of heart disease. In the image, the proportion of people with "obese weight" is significantly higher than that of "normal weight" and "over weight"

SleepTime

In the study of sleep time, we still refer to the reference given by the CDC. CDC defines," Most adults need 7 to 9 hours, although some people may need as few as 6 hours or as many as 10 hours of sleep each day. Older adults (ages 65 and older) need 7-8 hours of sleep each day " [4].



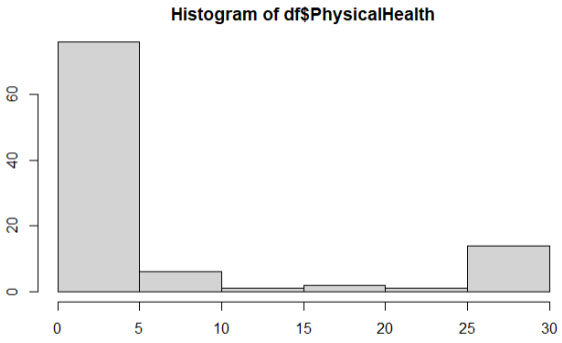
By observing the Sleeptime in our research data, we define less than 6 hours of sleep as "less sleep", 6-9 hours of sleep as "good sleep", and more than 9 hours of sleep as " over sleep". After processing, there are 35 objects were less sleep, 63 in good sleep, only 2 objects over sleep.

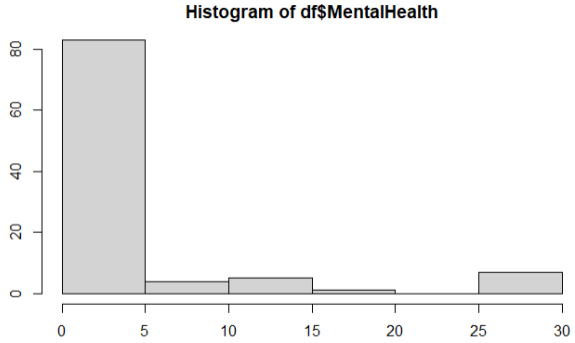
(Histgram of Sleep time)

PhysicalHealth and MentalHealth

There is similar between the data in PhysicalHealth and MentalHealth in histogram observation, both columns show a majority of subjects as 0 (never ill), and fewer people ever had ill. In contrast, there are more people who feel unwell every day. For example, we define 0 as "no Physical/Mental problem", 0-25 as "sometimes have Physical/Mental problem", 25-30 as "aways have Physical/Mental problem". Finally, close to 15% of the objects always have physical problems.

(Histogram of Physical/Mental problem)

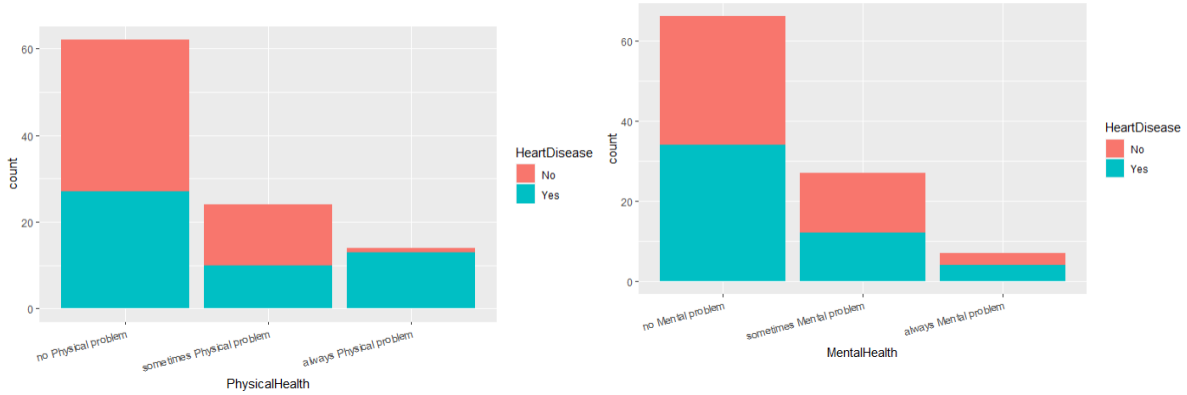


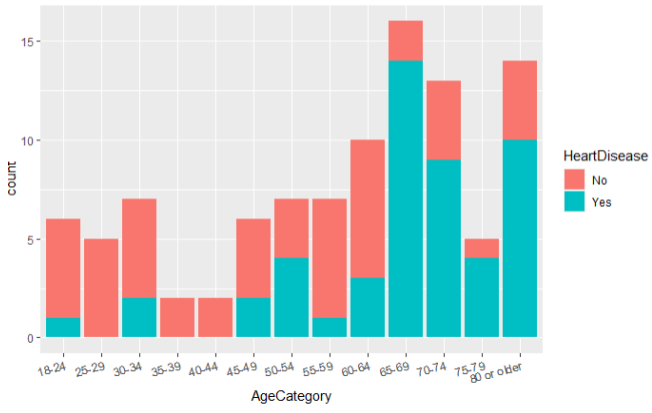


(Histogram of Physical/Mental problem with after classified value)

The study on the reclassification of PhysicalHealth and MentalHealth found that PhysicalHealth may be one of the factors affecting the disease. The main reason is that we found that almost all the people we classified as "always have Physical problem" had heart disease.

When studying other classes in the dataset, we found some possible contributors to the disease.

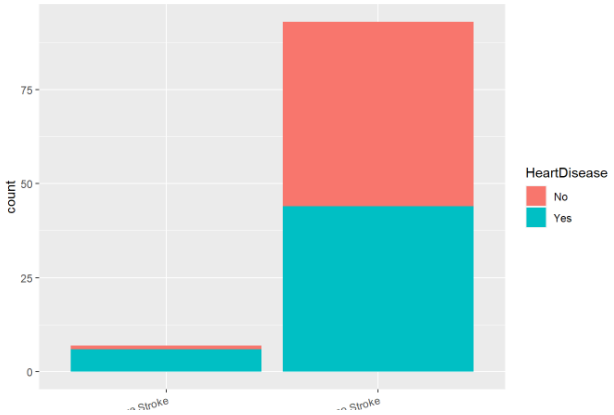


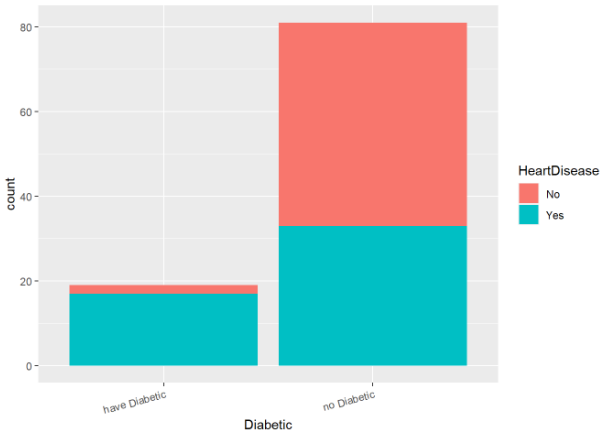


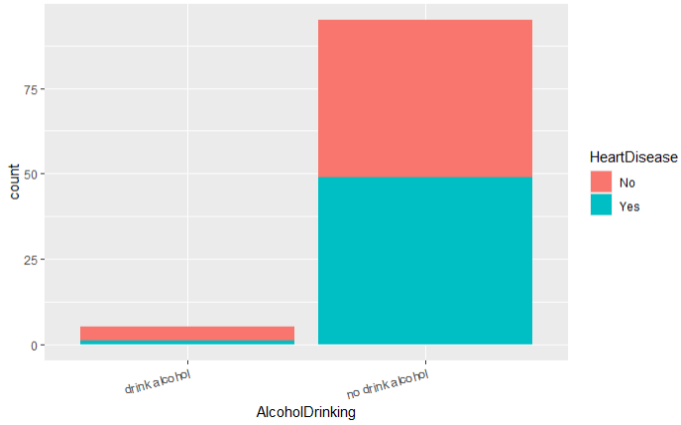
AgeCategory

Age appears to increase the likelihood of developing the disease. We analyzed the age and found that as the age increases, the proportion of the number of people who suffer from the disease is also gradually increasing. In the data set, most of the research subjects are over 60 years old, people around 65-69 age range in our data set with the largest, and people over 65 years old basically suffer from heart disease.

What goes against conventional thinking is drinking. It is common sense to think that alcohol has negative effects on the body. But in the data we studied there was no direct link between alcohol and heart disease. But this requires further verification because our data set does not represent the overall situation.







On the other hand, the data we randomly sampled may be too small, resulting in inaccurate information, because no correlation can be found in the study of the relationship between alcohol consumption and kidney disease. In fact, alcohol has certain effects on the kidneys, alcohol causes changes in the function of the kidneys and makes them less able to filter the blood [5]. Diabetes and stroke may be one of the factors that lead to the disease. In our study, patients with diabetes or stroke have a higher probability of also suffering from heart disease

Rules

In order to more efficiently explore the relationship between heart disease and different parameters in the data, we constructed LHS and RHS.

First in the LHS, we set the parameters as supp is 0.05 and conf is 0.8. With reference to our dataset, we found that the disease appears to be less likely in younger adults. We found that in the age range of 25-29, no one had heart disease, diabetes, stroke, kidney disease, skin cancer, etc.

	lhs <chr>	<chr>	rhs <chr>
[1]	{AgeCategory=25-29}	=>	{HeartDisease=No}
[2]	{AgeCategory=25-29}	=>	{Smoking=not smoke}
[3]	{AgeCategory=25-29}	=>	{DiffWalking=Ok to walk}
[4]	{AgeCategory=25-29}	=>	{Diabetic=no Diabetic}
[5]	{AgeCategory=25-29}	=>	{SkinCancer=no SkinCancer}
[6]	{AgeCategory=25-29}	=>	{Stroke=no Stroke}
[7]	{AgeCategory=25-29}	=>	{KidneyDisease=no KidneyDisease}
[8]	{AgeCategory=25-29}	=>	{AlcoholDrinking=no drink alcohol}
[9]	{AlcoholDrinking=drink alcohol}	=>	{PhysicalActivity=have PhysicalActivity}
[10]	{AlcoholDrinking=drink alcohol}	=>	{DiffWalking=Ok to walk}

In the RHS, we mainly explore which values are likely to cause heart disease. We changed the supp parameter to 0.04 to get more valuable information. And explore potential rules by sorting confidence, support, and lift respectively. We found that diabetes was the main cause, as we had previously suspected. Not only that, we speculate from the RHS that smoking and kidney disease are also the main causes of heart disease

lhs <chr>	<chr>	rhs <chr>
{Smoking=smoke, AlcoholDrinking=no drink alcohol, Sex=Male}	=>	{HeartDisease=Yes}
{Diabetic=have Diabetic}	=>	{HeartDisease=Yes}
{AlcoholDrinking=no drink alcohol, Diabetic=have Diabetic}	=>	{HeartDisease=Yes}
{Smoking=smoke, AlcoholDrinking=no drink alcohol, Sex=Male, Asthma=no Asthma}	=>	{HeartDisease=Yes}
{Smoking=smoke, AlcoholDrinking=no drink alcohol, Sex=Male, KidneyDisease=no KidneyDisease}	=>	{HeartDisease=Yes}
{Smoking=smoke, AlcoholDrinking=no drink alcohol, Sex=Male, Race=White}	=>	{HeartDisease=Yes}
{Diabetic=have Diabetic, Asthma=no Asthma}	=>	{HeartDisease=Yes}
{AlcoholDrinking=no drink alcohol, Diabetic=have Diabetic, Asthma=no Asthma}	=>	{HeartDisease=Yes}
{Smoking=smoke, AlcoholDrinking=no drink alcohol, Sex=Male, SkinCancer=no SkinCancer}	=>	{HeartDisease=Yes}
{Smoking=smoke, Sex=Male, Race=White}	=>	{HeartDisease=Yes}

Finally, we briefly studied "healthy" populations. We set the target heart disease=no through RHS, and we found several factors that may keep the heart healthy, including younger age, no smoking, no walking impairment, no diabetes

lhs <chr>	<chr>	rhs <chr>
{AgeCategory=25-29}	=>	{HeartDisease=No}
{AgeCategory=25-29, GenHealth=Good}	=>	{HeartDisease=No}
{Smoking=not smoke, AgeCategory=25-29}	=>	{HeartDisease=No}
{DiffWalking=Ok to walk, AgeCategory=25-29}	=>	{HeartDisease=No}
{AgeCategory=25-29, Diabetic=no Diabetic}	=>	{HeartDisease=No}

Machine Learning

Considering that our dataset is small, we plan to use 10-fold cross-validation with machine learning to build a predictive model. We first convert the data type, and then establish a training set and a test set according to 8:2.

First we use random forest classification and the average accuracy given by the training set is 0.75. In the process of cross-validation, the optimal parameters have been selected. We test in the test set, and the test accuracy is still 0.75.

(Random Forest)

```
          Reference
Prediction No Yes
       No   9   4
       Yes  1   6

               Accuracy : 0.75
                 95% CI : (0.509, 0.9134)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.02069
```

(SVM)

```
          Reference
Prediction No Yes
       No   7   3
       Yes  3   7

               Accuracy : 0.7
                 95% CI : (0.4572, 0.8811)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.05766
```

Secondly, we tried to use SVM. The average accuracy in the training set reached 0.79, but the accuracy in the test set was 0.7.

Improve

We want to try to improve the accuracy of model predictions. After previous data analysis, we found that some cloumns may have inaccurate information, or the proportion of Yes/No is too large. Therefore, we deleted 4 columns, including AlcoholDrinking, MentalHealth, Race, skinCancer. At the same time, we created a training set and a test set according to 8:2 from the new data. And re-use Randomforest and SVM to make predictions. From the conclusion, the average accuracy given in the training set has a 1~3% improvement. However, the accuracy presented by the Randomforest test set has not improved. In the new SVM model, the accuracy of the test set has increased by 0. 05. I think this is because our test sets too small.

(re-build Random Forest)

```
          Reference
Prediction No Yes
       No   9   4
       Yes  1   6

               Accuracy : 0.75
                 95% CI : (0.509, 0.9134)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.02069
```

(re-build SVM)

```
          Reference
Prediction No Yes
       No   9   4
       Yes  1   6

               Accuracy : 0.75
                 95% CI : (0.509, 0.9134)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.02069
```

Conclusion

Firstly, through image and RHS analysis, we speculate that several causes of heart disease are advanced age, high BMI, smoking, and diabetes.

At the same time, we built RandomForest and SVM models to predict whether our subjects may suffer from heart disease. From the analysis of the results, there is no significant difference in the prediction accuracy of SVM and RandomForest, and even the accuracy of RandomForest is higher. But we think the SVM model works better.

In the confusion matrix of observing the test results of RandomForest and SVM, the value of SVM in False positive is greater than that of RandomForest. At the same time, among the 20 subjects tested, SVM judged that 10 people were sick, while RandomForest only judged that 7 people were sick. This shows that SVM is more inclined to give False positive results. In this heart disease case we are studying, we believe that SVM can give an early warning of the tester's physical condition and tell the tester that the body is in an unhealthy state in time. So, from a certain point of view, our SVM model is very successful.

Citations

[1] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[2] https://www.cdc.gov/heartdisease/about.htm

[3] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/

[4] https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html

[5] https://www.kidney.org/news/kidneyCare/winter10/AlcoholAffects#:~:text=Alcohol%20causes%20changes%20in%20the,and%20organs%2C%20including%20the%20kidneys.