Study on Key Personal Indicators of Heart Disease

Multivariate final Project

庄成 2019080344

**Abstract**:

According to the CDC, heart disease is one of the leading causes of death for people of most races in the United States. The CDC have listed out 18 key personal indicators of potential Heart Problem. In this study, we aim to narrow down the 18-dim variable space into a lower dimension space while interpreting the numerical and real-life significance of the results. Furthermore, we would also try and build up a mathematic model that can be used to predict the potential threat of a heart problem from the result of a given data input.

**Keywords: Heart Disease, CDC, PCA, Fisher LDA, Logistic Regression**

Table of Content

1. Background and Significance1
2. Dataset Description4
3. **Dataset Pre-Processing1**
4. **Dimension Reduction1**
5. **Principle Component Analysis1**
6. **Factor Analysis1**
7. **Modelling1**
8. **Discussion, Limitations, and Improvements1**
9. **Conclusion1**
10. **Background and Significance:**

According to the CDC, heart disease is one of the leading causes of death for people of most races in the United States. This includes Caucasian, African American, Hispanic, and American Indians. The CDC has gone on with great effort to put together a list of key factors for potential heart disease, including high blood pressure, high cholesterol, smoking, diabetic status, obesity, lack of physical activities, and alcohol abuse. They then interviewed and collect 30,000+ patients with or without heart disease histories and the status of the key indicators.

The human heart is the center of all body functions, it is the most vital organ within the human biological system. Cardiac diseases can lead to irreversible and perpetual health damage. Therefore, it is of the paramount importance to have the capability of examining and diagnosing patients before they actually experience the devastating trauma. Moreover, with the weighted importance of the indicators, people will have a better understanding on what measurement they should be taking, thus a higher chance of preventing the risk of potential heart diseases.

In this study, I will mainly address the following three questions:

1. Is there any determining factor amongst all the listed key indicators?
2. Are we able to narrow down the list and eliminate the less important indicators?
3. The prediction of the results of potential heart disease based on the indicators provided
4. **Dataset Description:**

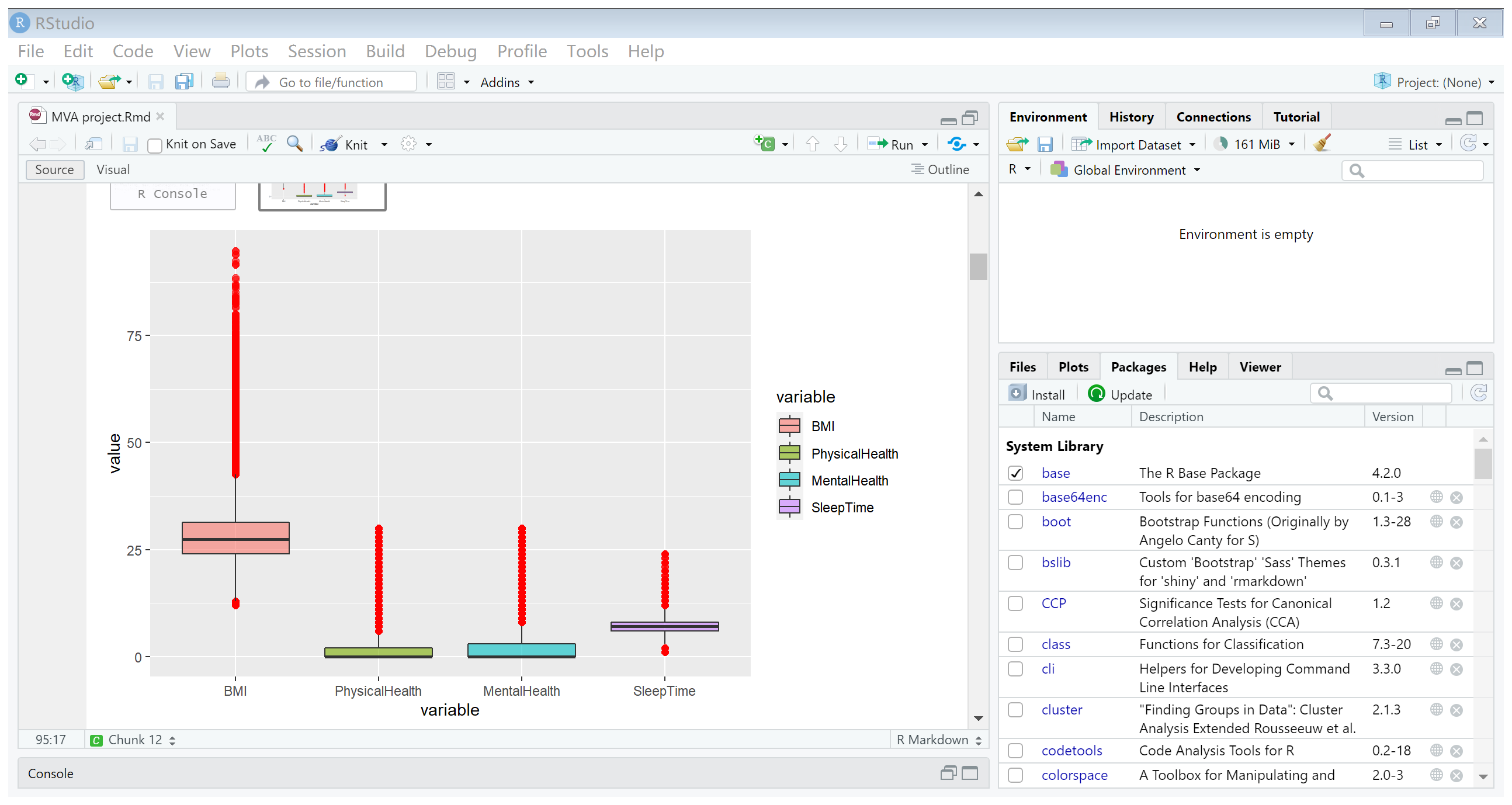
The original dataset is downloaded from Kaggle, with the name “Personal Key Indicators of Heart Disease”. The original dataset is presented by BRFSS, which consists of over 400,000 adult interviews and over 279 variables mostly made up of questions. The Author of the Kaggle dataset narrow-downed the variables by relevance and qualified all the inputs, making it easier for further data processing. The variables are listed in the following table:

|  |  |  |
| --- | --- | --- |
| Variable | Qualification | Data Type |
| **HeartDisease** | Reported Coronary heart disease? | Boolean |
| **BMI** | Body Mass Index | Int |
| **Smoking** | Smoked more than 100 cigarettes ？ | Boolean |
| **AlcoholDrinking** | adult men having more than 14 drinks per week and adult women having more than 7 drinks per week | Boolean |
| **Stroke** | Stroke History? | Boolean |
| **PhysicalHealth** | Number of days of physical illness in the past 30 days | Int |
| **MentalHealth** | Number of days of mental illness in the past 30 days | Int |
| **DiffWalking** | Serious difficulties in walking or climbing stairs | Boolean |
| **Sex** | Male/Female | Char |
| **AgeCategory** | Age level with categories: 18-24; 25-29; 30-34; 35-39; 40-44; 45-40; 50-54; 55-50; 60-64; 65-70; 70-74; 75-79; 80 or older | Categories |
| **Race** | Ethnicity: White, Hispanic, Black, Asian, American Indian & Alaskan Natives, Others | Char |
| **Diabetic** | Diabetes history? | Boolean |
| **PhysicalActivity** | Had done physical activities in the past 30 days | Boolean |
| **GenHealth** | General self-evaluation: Poor, Fair, Good, Very Good, Excellent | Char |
| **SleepTime** | Hour of sleep per day | Int |
| **Asthma** | Asthma History | Boolean |
| **KidneyDisease** | Kidney disease history? (except kidney stones, bladder infection or incontinence) | Boolean |
| **SkinCancer** | Skin cancer history? | Boolean |

*Table 1: Summary of dataset variables*

*Figure 1: Example of the dataset*

As shown in the graph above, most of the data are binomial inputs of Yes and No and descriptive data. However, there are 5 continuous data types, and its data summary are shown below:

****

*Figure 2: The description of non-binary variables*

The are a couple of outlier within each variable, which is helpful in analyzing anomalies.

1. **Dataset Pre-Processing**

|  |  |
| --- | --- |
| Variable | Category |
| **Sex** | Male = 1; Female = 0 |
| **AgeCategory** | 18-24: 1; 25-29: 2; 30-34: 3; 35-39: 4; 40-44: 5; 45-40: 6; 50-54: 7; 55-50: 8; 60-64: 9; 65-70: 10; 70-74: 11; 75-79: 12; 80 or older: 13 |
| **Race** | White = 1; Black = 2; Asian = 3; Hispanic = 4; Native American/Alaskan Natives = 5; Others = 6 |
| **Diabetic** | Yes = 1; No = 0; Borderline = 1.5; During Pregnancy = 2 |
| **GenHealth** | Poor = 1; Fair = 2; Good = 3; Very Good = 4; Excellent = 5 |
| **Diabetic** | Yes = 1; No = 0; Borderline = 0.5; During Pregnancy = 2 |

Table : Legend of different variables

As seen in the dataset description, there are a lot of Boolean type variables and a couple of level categorized variable. These characteristics will intervene with later on data analysis. Therefore, for the sake of convenient and analysis purposes, I will convert all Yes and No discreate data into 1 and 0 and all level categorized data into a ranked level started with 1 for the first category, 2 for the second category, and so on. The legend is shown is the table below:

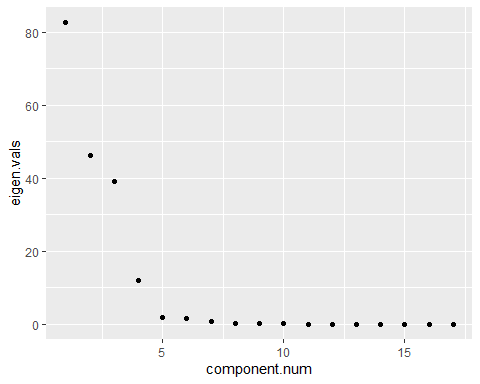
*Table 2: Legend of different variables*

The data after processing looks like as followed:

*Figure 3: Data after transformation*

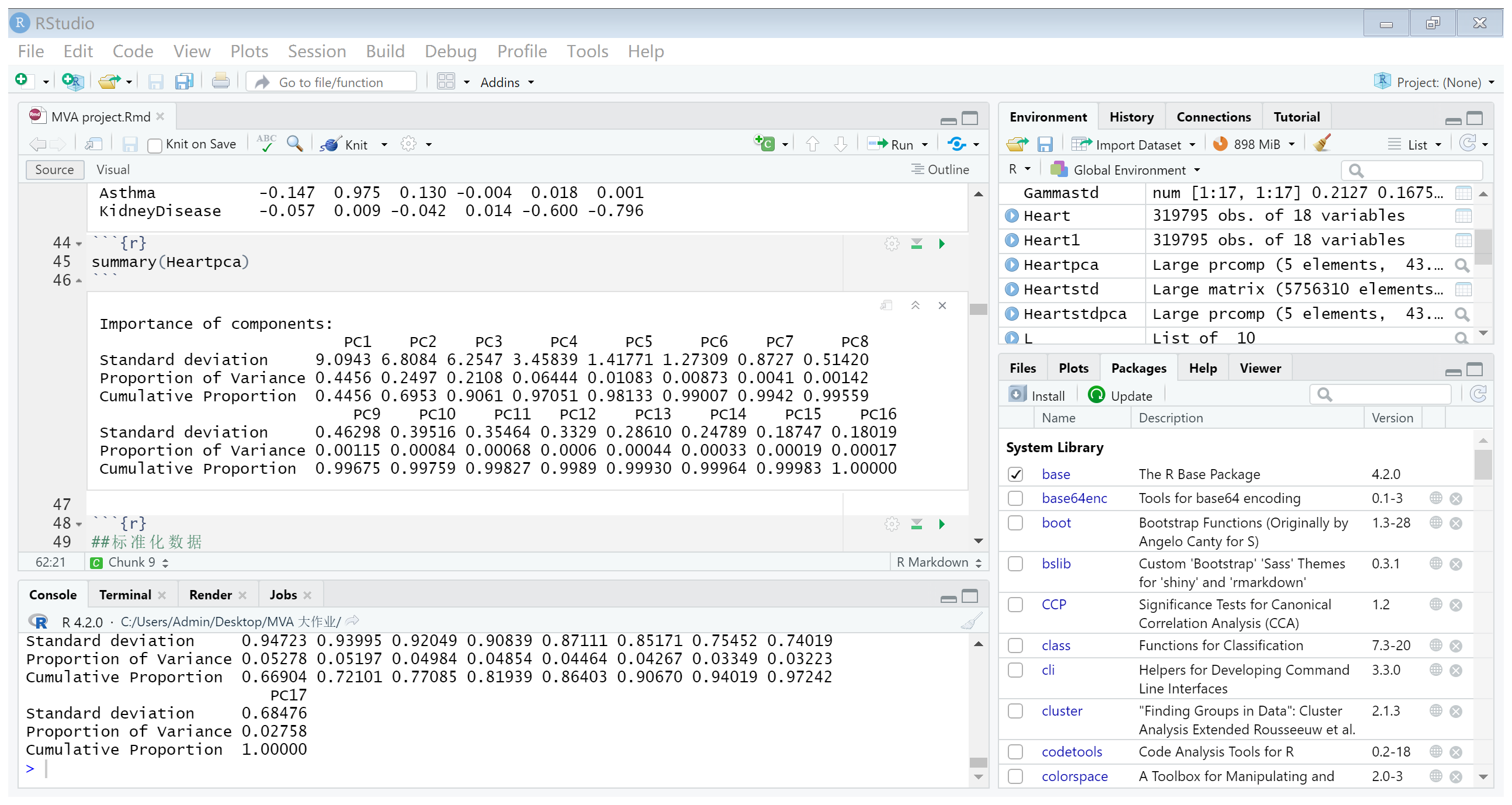
1. **Dimension Reduction**
   1. **Principle Component Analysis**

This heart disease dataset consists of 18 key indicators in total, it is not very helpful to the patients when they want to self-diagnose themselves with so many factors. Furthermore, the 18 key indicators are somewhat related, for example if a patient does a lot of physical activities, they are probably in a good physical health state, that means they have a better BMI score, better general health, and they most likely don’t experience any difficulty in walking. Therefore, we want to try to reduce the dimension of the key indicators, we aim to see if we can use fewer key indicators to explain the majority of the dataset. This will help us better understand the true origins of heart disease. In order to achieve this goal, we utilize the method of Principal Component Analysis (PCA) in multivariate statistics, and the results are shown below:



*Figure 4: Elbow graph of PCA*

From the graph above, we can tell that there is a significant turning point at component num. = 5. The eigenvalue from 5 and onwards are small ad stagnant. This means that most of the dataset can already be described by the first 5 eigenvalue and eigenvector.

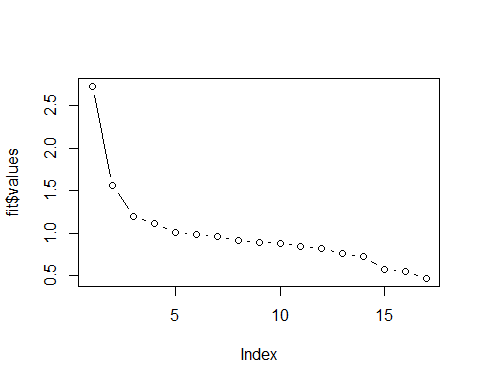
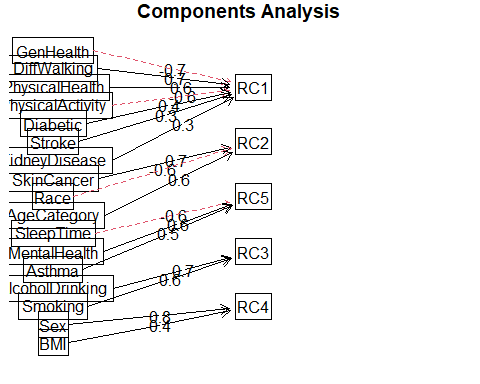


*Figure 5: Results of PCA*

Let’s take a glance at the numerical importance and coverage of the PCs. We can see that when reaching 5 principal components, 98.13% of the data in the original dataset can already be explained. Therefore, a 5-principal component dimension model is indeed appropriate.

* 1. **Factor Analysis**

The effect of the of 5 principal component dimension reduction in the PCA model can be further augmented by the Factor Analysis (FA) model. What FA achieves is that it tries to retain as much original information as possible while maximizing the difference between each variable. There are two ways of performing the Factor Analysis, the PC approach and the MLE approach. However, the MLE approach is only conducted under the normality assumption, which in the Heart dataset, is not applicable. Therefore, we would be taking the PC approach. The results of FA are as followed:

*Figure 6: Elbow graph of FA Figure 7: FA diagram*

From the diagrams above, we can also see a slight turning point around the 5th index point, thus we can safely conclude that 5-principal component dimension is an appropriate choice. With the other objective in mind, we want each of the key indicator to be mainly explained by one latent factor. From the graph on the right we can categorize the indicators in to the following latent factors. RC1 mainly affects physical state and physical wellbeing of a patient, including: GenHealth, DiffWalking, PhysicalHealth, PhysicalActivity, Diabetic, Stroke, Kidney Disease. RC2 mainly affects the demography of the patient, including Race, AgeCategory, and Skin Cancer. Maybe there is a correlation between skin cancer and people’s race and age. RC3 mainly affects patient’s lifestyle, including Smoking and AlcoholDrinking. RC4 affects the Sex and BMI, this however baffles me because I cannot see a clear connection between the two variables. Last, RC5 mainly affects the mental wellbeing of a patient, including SleepTime, MentalHealth, and Asthma. This furthers the explanation of the reduced dimension space.

1. **Grouping**

In this dataset, we are presented with over 30,000+ data inputs and with only two outcomes, either yes or no to having heart problems. With so many data, we want to experimental trial to use just the indicators to group the data inputs into the ‘Yes’ and ‘No’ result categories. This will help us with predicting the patient’s potential threat to heart related diseases.

The method we can use is the Fisher’s Linear Discrimination Analysis (LDA). The entire idea of Fisher’s LDA is to find a vector , that projects all the samples on to a new axis , which, on Y-axis, maximizes the difference between the mean of the two groups and minimize the scatter of the two groups. The result is as followed:

After obtaining and the linear discriminate model , we need to test of the accuracy of the model. In order to avoid over-fitting, we made use of cross validation in the LDA model and the Apparent Error Rate (APER) is chosen to measure the accuracy of the model. The results are as followed:

|  |  |  |
| --- | --- | --- |
| True  Pred | 0 | 1 |
| 0 | 284,645 | 21,395 |
| 1 | 7,777 | 5,978 |

*Table 3: LDA prediction results*

The overall the APER of the result is 91.22%, which is looks acceptable. However, if you calculate the accuracy separately, it becomes less acceptable. The table shows that out of the 292,422 patients who doesn’t have heart problem, 284,645 were grouped correctly and 7,777 mistakes were made. The accuracy is around 97.3%. On the other hand, out of the 27373 patients who have a heart problem, 21,395 were falsely grouped. The accuracy for is only at 21.8%, which is very low.

1. **Modelling**

The LDA method we used to predict the results have been less than satisfying, therefore, we want to test out another method and see if the prediction accuracy can be improved. We now turn to the regression model. Since this is a Discrete model with only two outcomes, we will use the logistic regression to model this dataset. The logistic regression model is as followed:

The results are as followed:

We can make a couple of inferences from the model, patients with stroke history, patients who smokes, patients with diabetes, and patients with kidney disease increases the chances of incurring a heart disease more significantly than other indicators, while patients with a good general health status decrease the chances more significantly than others.

Although we came to some conclusions with the help of this model, we still want to test out whether the model is a good prediction of the data or not. In order to attest the real accuracy of the model, we separated the dataset into a training set and a prediction set and predict the results on the prediction set. The outcomes are as followed:

|  |  |  |
| --- | --- | --- |
| True  Pred | 0 | 1 |
| 0 | 17,779 | 1,652 |
| 1 | 168 | 196 |

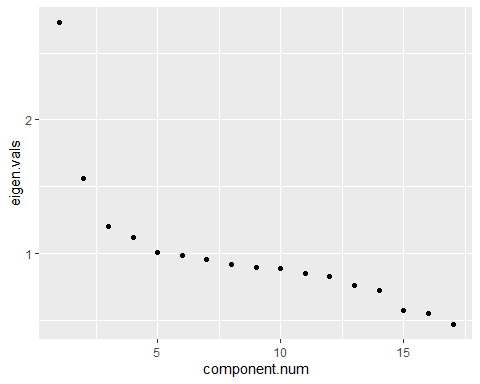
*Table 4: Regression Model prediction result*

The error rate for healthy patient has reached a new low of 0.9%, however the error rate of heart problem patients has also reached a new high of 89.4%. There are still great room for improvement, and the reason behind this phenomenon will be discussed in the next section.

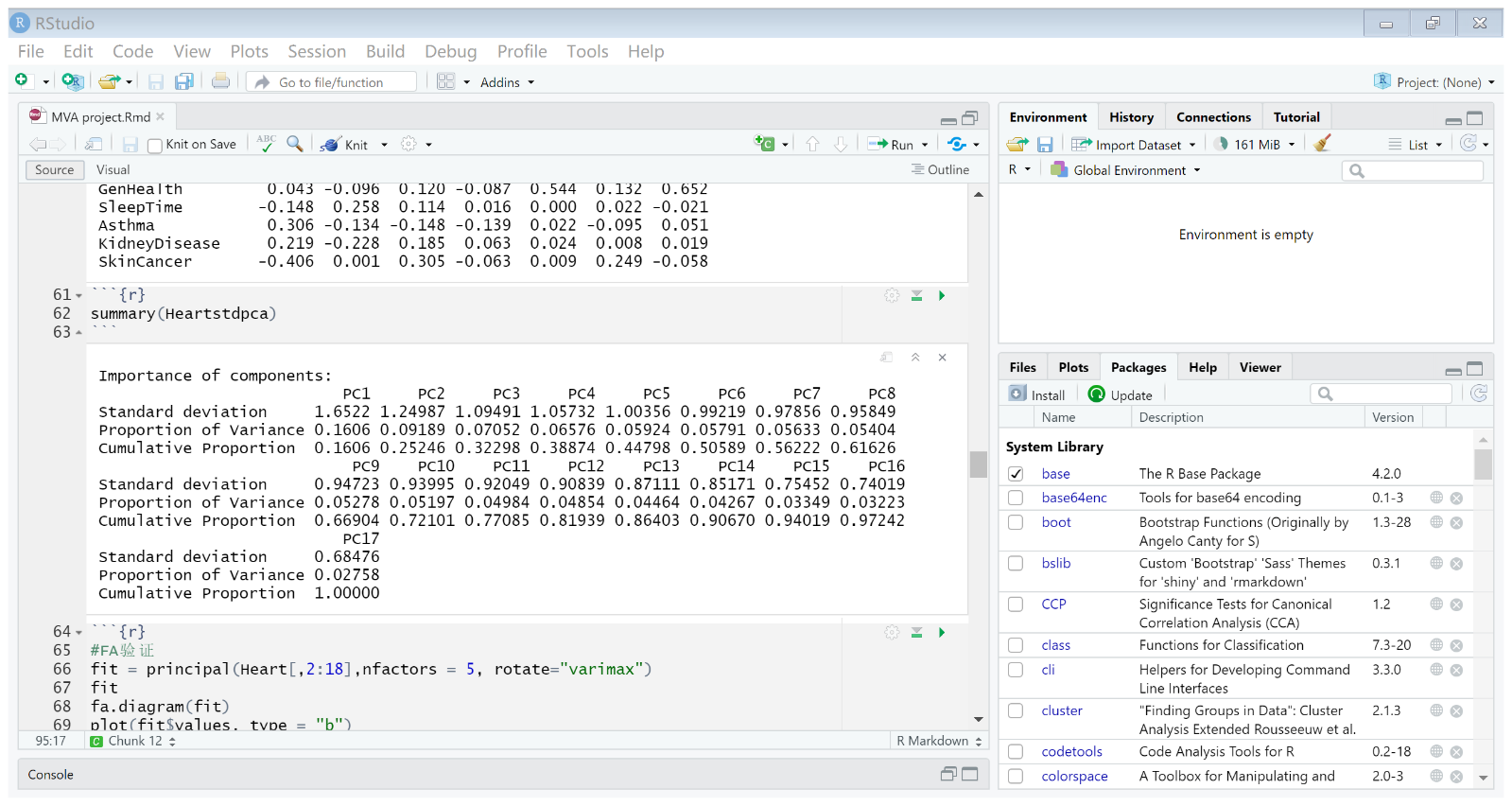
1. **Discussion, Limitations, and Improvements**

There are a few interesting occurrences and limitations during the compiling of this project which I want to point out and explain.

1. The standardization of the original dataset.



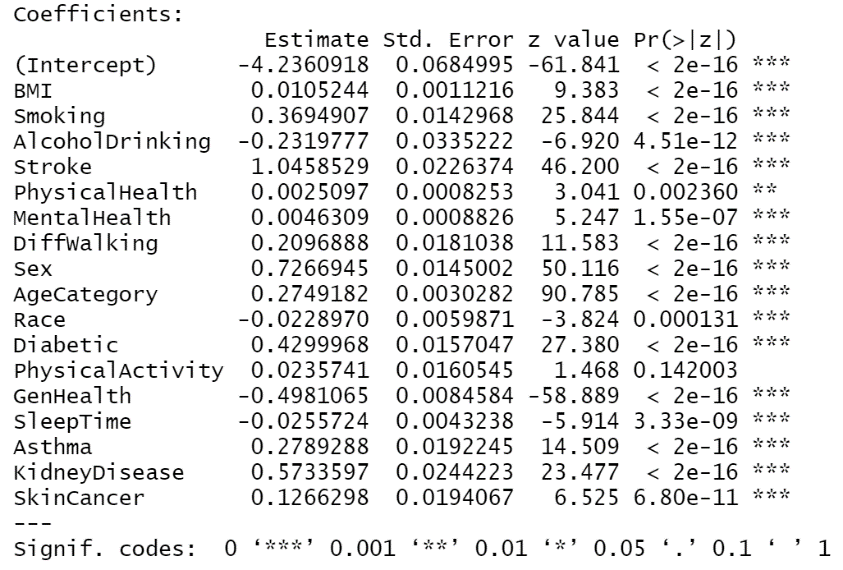
*Figure 8: Elbow Graph of Standardized PCA*



*Figure 9: Results from Standardized PCA*

These are the two results of the PCA of the standardized Heart Disease Dataset. When compared to the unstandardized dataset, the dimension reduction efficiency has significantly decreased, and its cumulative proportion of the dataset explained have also decreased significantly. My speculation of the reason behind such phenomenon is that all variables have different level of significance before the standardization. When standardized, all variables are ignored of their real-life importance and scaled from a united point of view. This is quite contradicting with the personal indicators because they were meant to hold different importance. Thus, the standardization eventually led to the need of more dimensions to explain the dataset.

1. The importance of the variables in the logistic regression model.



*Figure 10: Results of Regression Model*

As you can see, 16 out of 18 variables have been rated with a ‘\*\*\*’ significant, meaning all 16 variables play an important role when predicting the outcome. This is because more variables are binomial inputs, either 1 or 0. Thus, a single change in input alters the f(x) function and thus the outcome of the prediction model.

1. The prediction accuracy of the regression model.

The results of the regression model are less than acceptable. The model is very biased in terms of prediction. The model can successfully predict the healthy data, but not the actual heart problem patient. This is mainly due to the lack of heart patient data, there are only 27,000 heart problem patient data out of the 300,000+ data inputs. This has seriously affected the training of the regression model. Improvements can be made by collecting more heart patient information into the dataset.

1. The lack of normality

The inputs of the dataset are all binomial, this has limited the choices of analysis tools that can be utilized. Improvements can be made by increasing the scales of each discrete binomial input, turning each input into a continuous normally distributed input. This would increase the efficiency of the analysis and deepen the interpretation of the entire problem.

1. **Conclusion**

During this study, we have found that the 18 personal key indicators can be categorized into 5 labels: physical wellness, demography, lifestyle, gender, and mental wellbeing. This means that these are the main factors that might affect the wellbeing of a person’ heart. Moreover, we were able identify the importance of the different indicators. Indicators such as stroke history, diabetes, kidney disease, smoking, and gender affects the heart the most. Meaning people should at all cost quit smoking and avoid stroke, diabetes, and kidney risk. Lastly, we were able to predict the results of a patient with a given input. Although the results are not as satisfying and still needs improvement, we did come to a few conclusions and we can further tackle this problem head-on.