가로선

**National Cancer Center Graduate School of Cancer Control and Population Health**

Statistical Computing Class

06.DEC.2021

2104202

Kwon.Hoejun

Identify the Heart Disease risk factor using BRFSS data : Validity evaluation of variables in predicting heart disease

# 1. Introduction

Heart Disease is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy.

The majority of individuals only learn they have the disease following symptoms such as chest pain, a heart attack, or sudden cardiac arrest. This fact highlights the importance of preventative measures and tests that can accurately predict heart disease in the population prior to negative outcomes like myocardial infarctions (heart attacks) taking place.

# 2. Task

**1. To what extend can survey responses from the BRFSS be used for predicting heart disease risk?**

**2. Can a subset of questions from the BRFSS be used for preventative health screening for diseases like heart disease?**

# 2.Materials and Methods

**A) Materials**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services.

Given these risk factors, I tried to select features (columns/questions) in the BRFSS related to these risk factors. To help understand what the columns mean, I consult the BRFSS 2015 Codebook to see the questions and information about the questions. I try to match the variable names in the codebook to the variable names in the dataset I downloaded from Kaggle. I also reference some of the same features chosen for a research paper by Zidian Xie et al for *Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques* using the 2014 BRFSS. Diabetes and Heart Disease outcomes are strongly correlated, with the primary cause of death for diabetics being heart disease complications.

BRFSS 2015 Codebook: <https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf>

BRFSS 2015 Data : <https://www.cdc.gov/brfss/annual_data/annual_data.htm>

Relevant Research Paper using BRFSS for Diabetes ML: <https://www.cdc.gov/pcd/issues/2019/19_0109.htm>

**B) Method**

1. **Logistic Regression**

1. Estimates each variables the impact through univariable regression

2. Consider multiple regression models except for variables with low influence.

3. The estimated variables may vary depending on the regression method, so Stepwise, forward, backward. Check the difference through all methods.

4. If the variables obtained through each method are the same, construct a reduced model through the obtained variables.(Because variables deleted through the regression process may exist.)

5. Check the total variables, and additionally separate them by gender to check the influence of each variable.

1. **Factor Analysis**

1. By grouping variables into a common dimension and reducing them to a small number of factors, an explainable relationship is derived.

2. Determine the number of factors and evaluate the extracted factors.

3. Rotate the factor and calculate the score.

4. Regression analysis is performed through the obtained factor scores to confirm the effect of factors on the response variable.

1. **Random Forest**

1. Creates an ensemble of hundreds of decision trees to predict a single target of either interval or nominal measurement level

2. Deletes from the training data any observation that has a missing target value or a FREQ variable whose value is less than or equal to 0.

3.Identify the relative neutrality of each variable by distinguishing between division-based and observation-based variable importance measurements.

# 3. Process

Check whether variables are missing from the data through primary preprocessing.

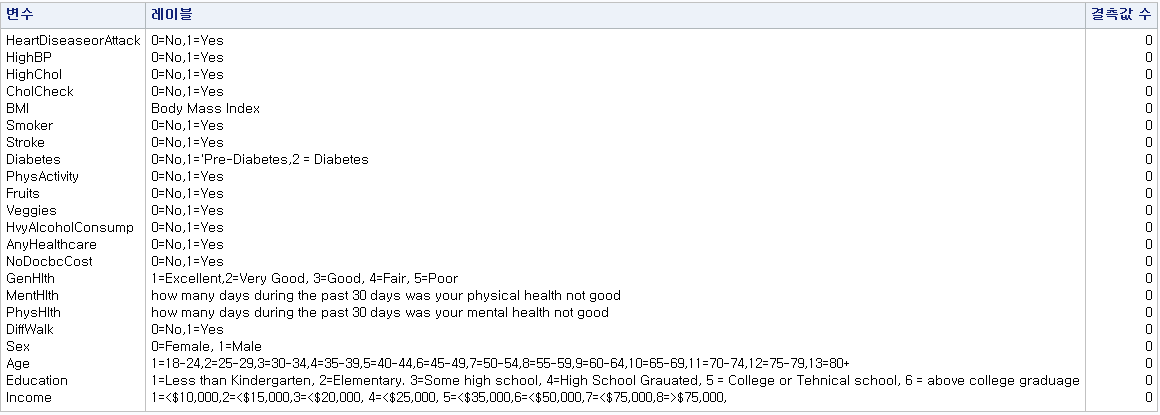


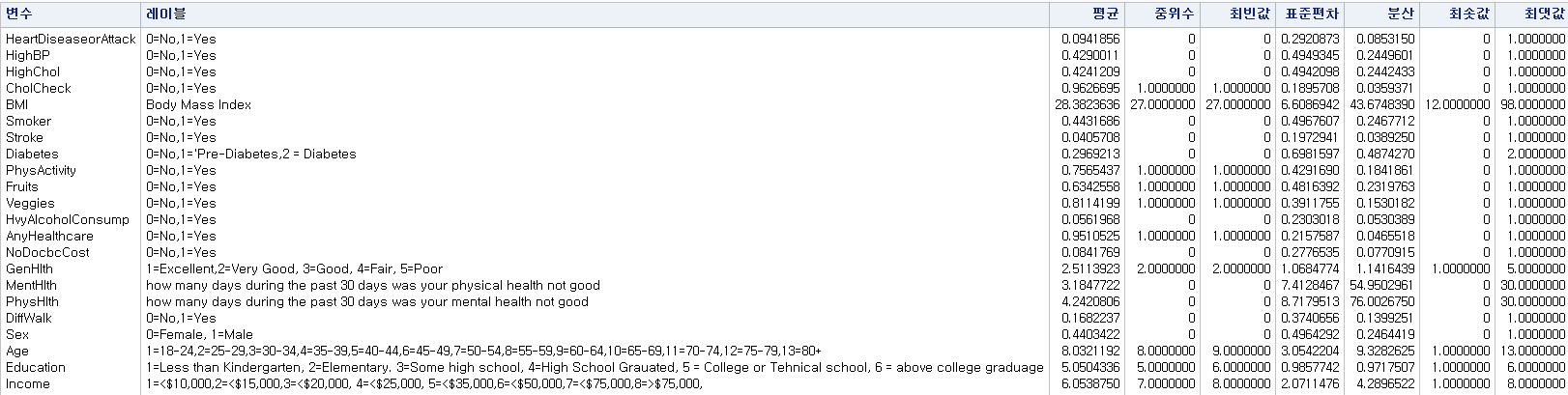
Labeling is carried out to facilitate variable identification.

When the data were checked, continuous or normal data was identified, not in the binary form.

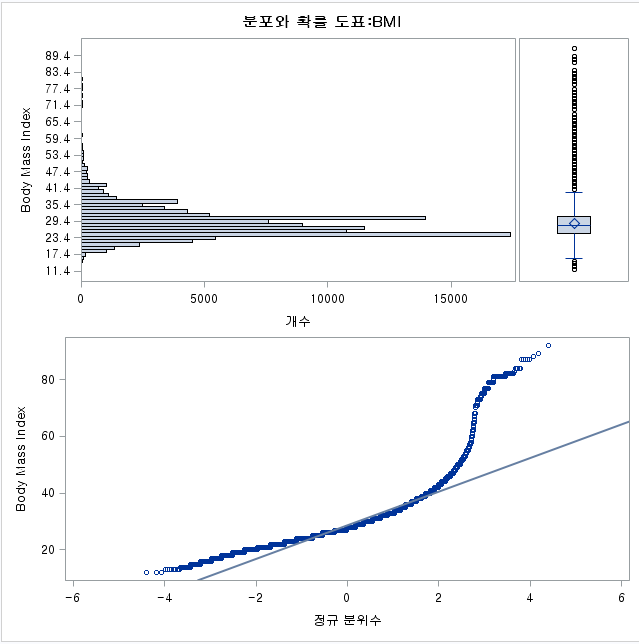
In the case of ordinary data or continuous data, this part needs to be considered because if the distribution of specific data is biased, there may be biased parts about the influence of variables.

Before checking for each variable in detail, check the missing value of the variable and the overall summary statistic to check the outlier or extreme value.



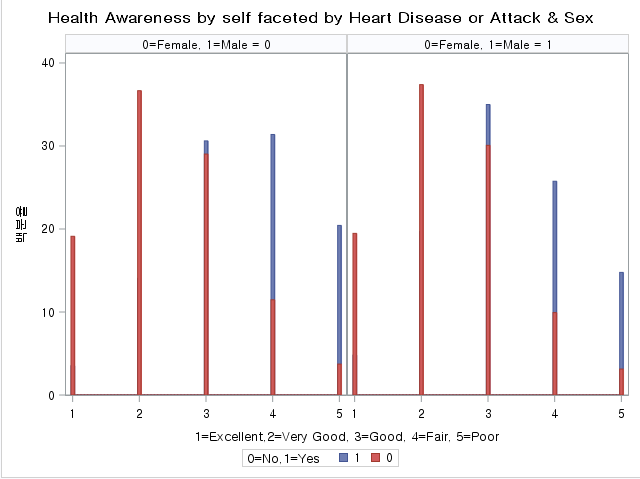


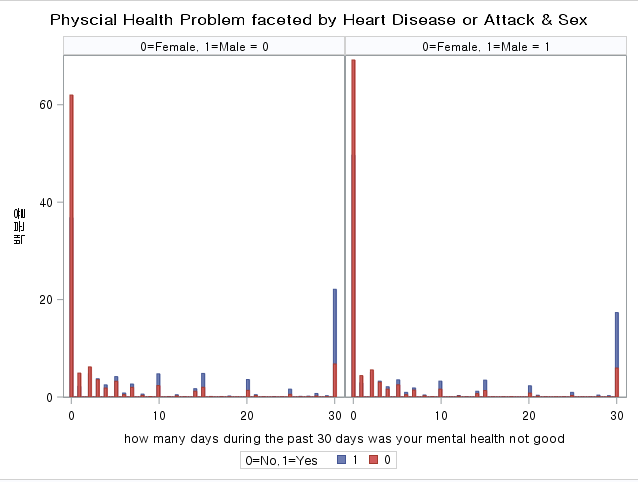
When I checked the missing values, it was confirmed that there was no problem in pretreatment. However, in summary statistics, there is a need to check the extreme value of BMI. When the BMI was the lowest price of 12 and the highest price of 92, it was necessary to check whether the data was generated due to the data notation error or whether the actual BMI was correct. If it is a simple data notation error, many data will not show such a value, so it was decided to check the overall distribution of BMI according to gender.



Due to a number of BMI values located at extreme values, it can be seen that the overall data violates the normality. In addition, a large number of BMIs of 50 or more are observed, and the values in the extreme values are also considered to be BMI of the actual subject, and accordingly, it was decided not to remove the variable by determining them as outliers. However, if such data is used as it is, a normality problem also occurs, so it was decided to convert it into ordinal data and proceed.

The next variable is General Health.

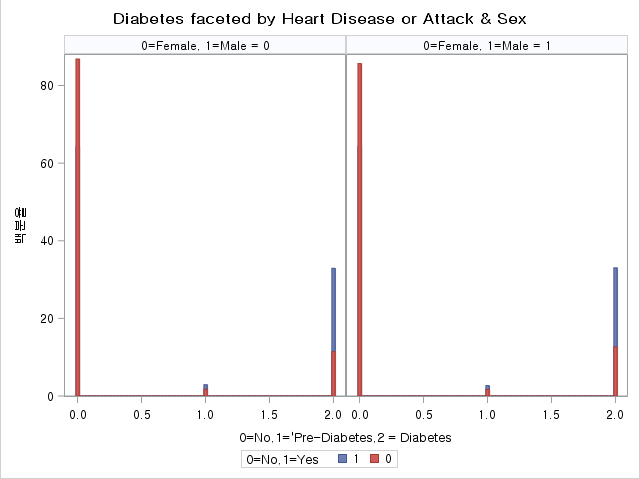


From 1 to 5, it is a measure of recognizing that the subject's health status has deteriorated, and since there is a difference in incidence according to perception according to gender, the data is used as it is, but the order of the scale is changed to facilitate understanding of variables.

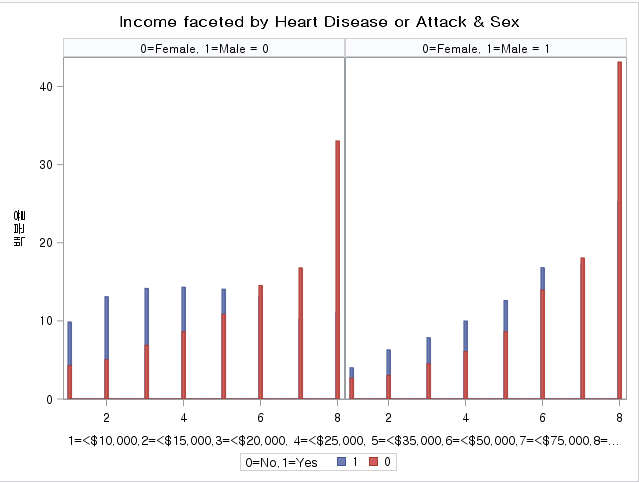
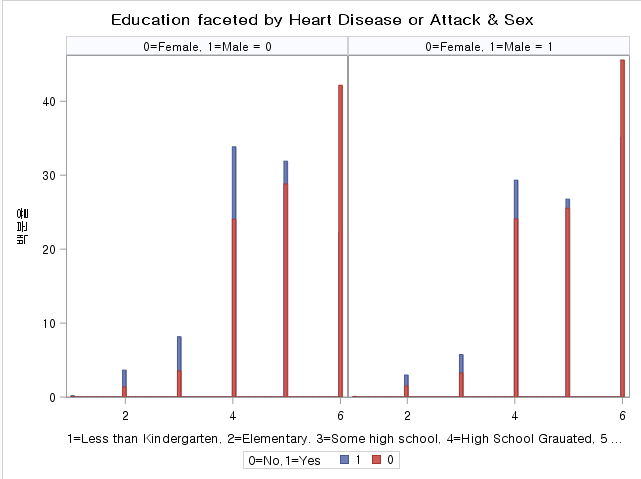
The second is a variable representing the scales of the Mental Health Problem and the Physical Health Problem, which means the period of discomfort over the past month.

It is important how long you have suffered from discomfort in a short period of time, such as the previous data using BRFSS, to confirm that there are various conditions that can simply temporarily constrain you.

Therefore, the degree of discomfort according to the period is recodingto 0 and 1.



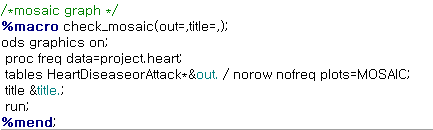
In the case of the Diabetes variable, it is separated into Pre-Diabetes and Diabetes, not simply the presence or absence of disease, and 90% of Pre-Diabetes develop into Diabetes, and the proportion of Pre-Diabetes is also small, so it is merged with Diabetes.

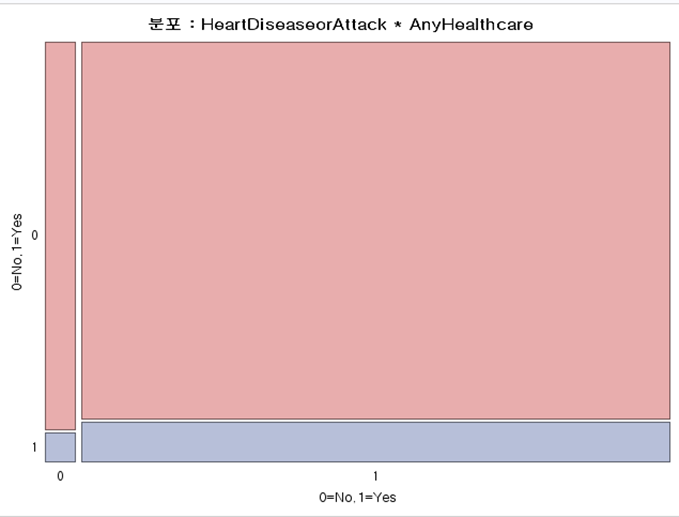
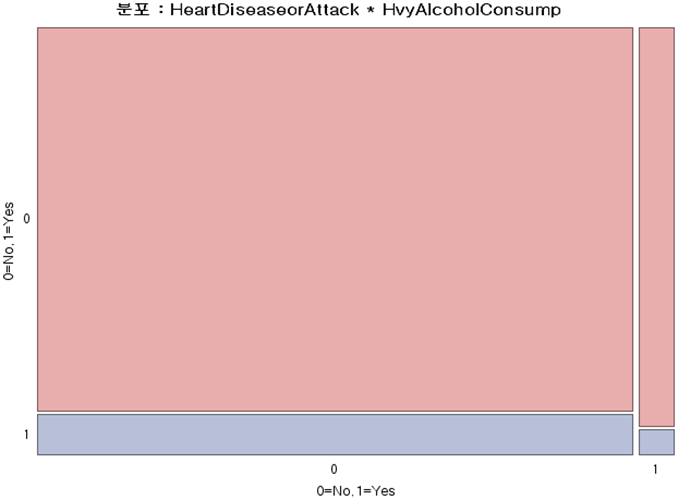


In the case of Education and Income, it can be seen that the proportion of heart disease is very small in the case of the highest educational level and the highest economic level. Even in this case, since there are no specific patterns for the original variables, and the frequency of each original data is different, it is re-coded by dividing into high-income and highly educated people to match the overall weight and construct easily interpretable variables. At this time, the standard for high-income earners is more than $75, 000, and those with high education are judged to be more than college students or technical schools.

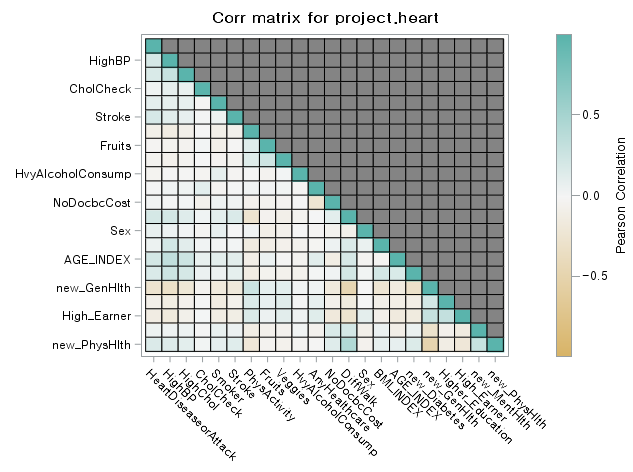
The code for re-coding is as above.

In order to concisely see the distribution of all variables, including the variables previously investigated, it is briefly displayed as a mosaic plot. Repeated processing was facilitated using macros.



Unlike other variables, heavy alcohol acceptance, cholesterol check, and AnyHealthCare variables tend to have a rather low incidence rate in people with negative tendencies.

In order to measure the correlation between variables, a heat map plot according to the score for each correlation was created and confirmed.

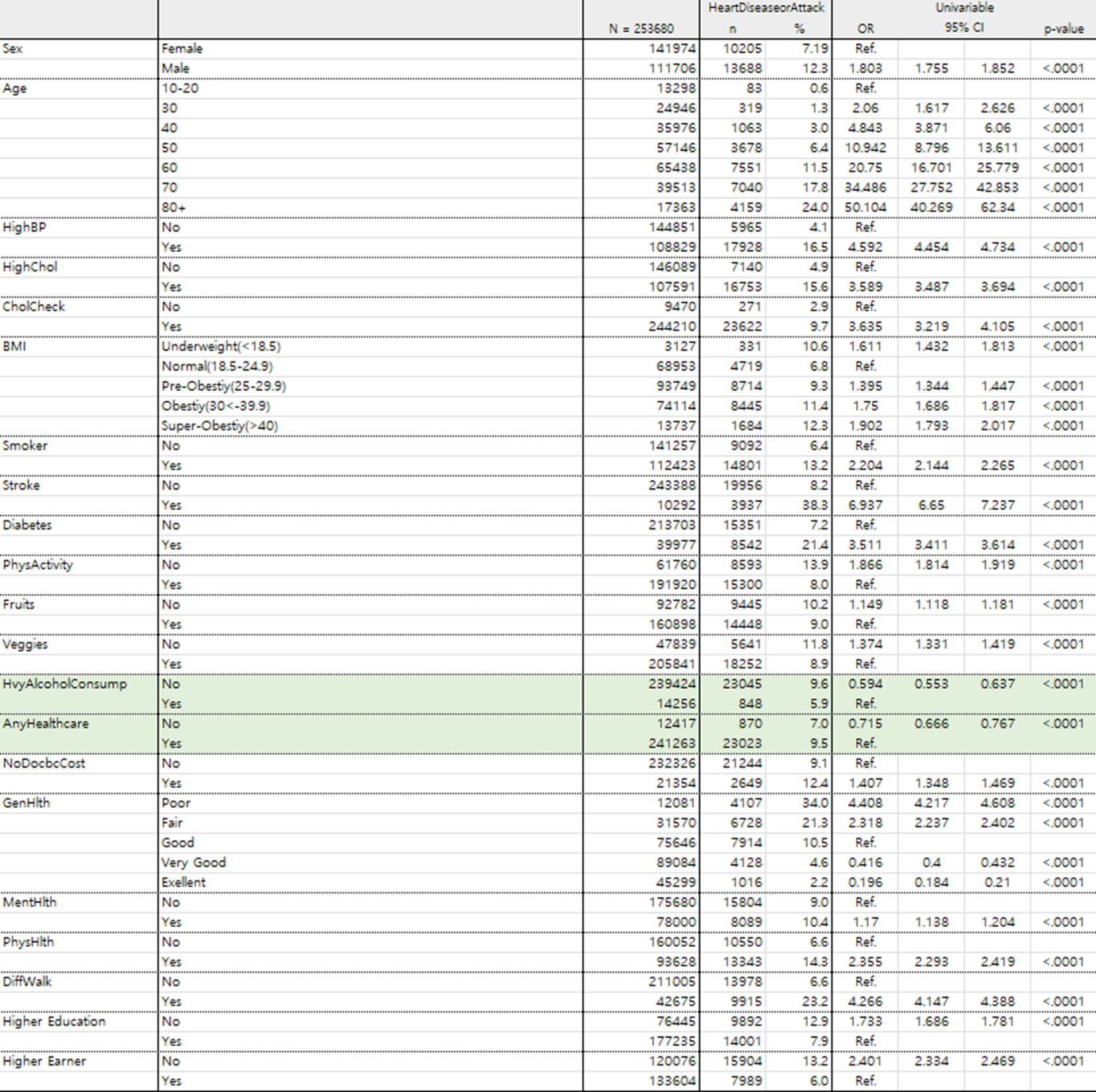


The above figure is a heatmap plot schematically illustrating the correlation between variables.

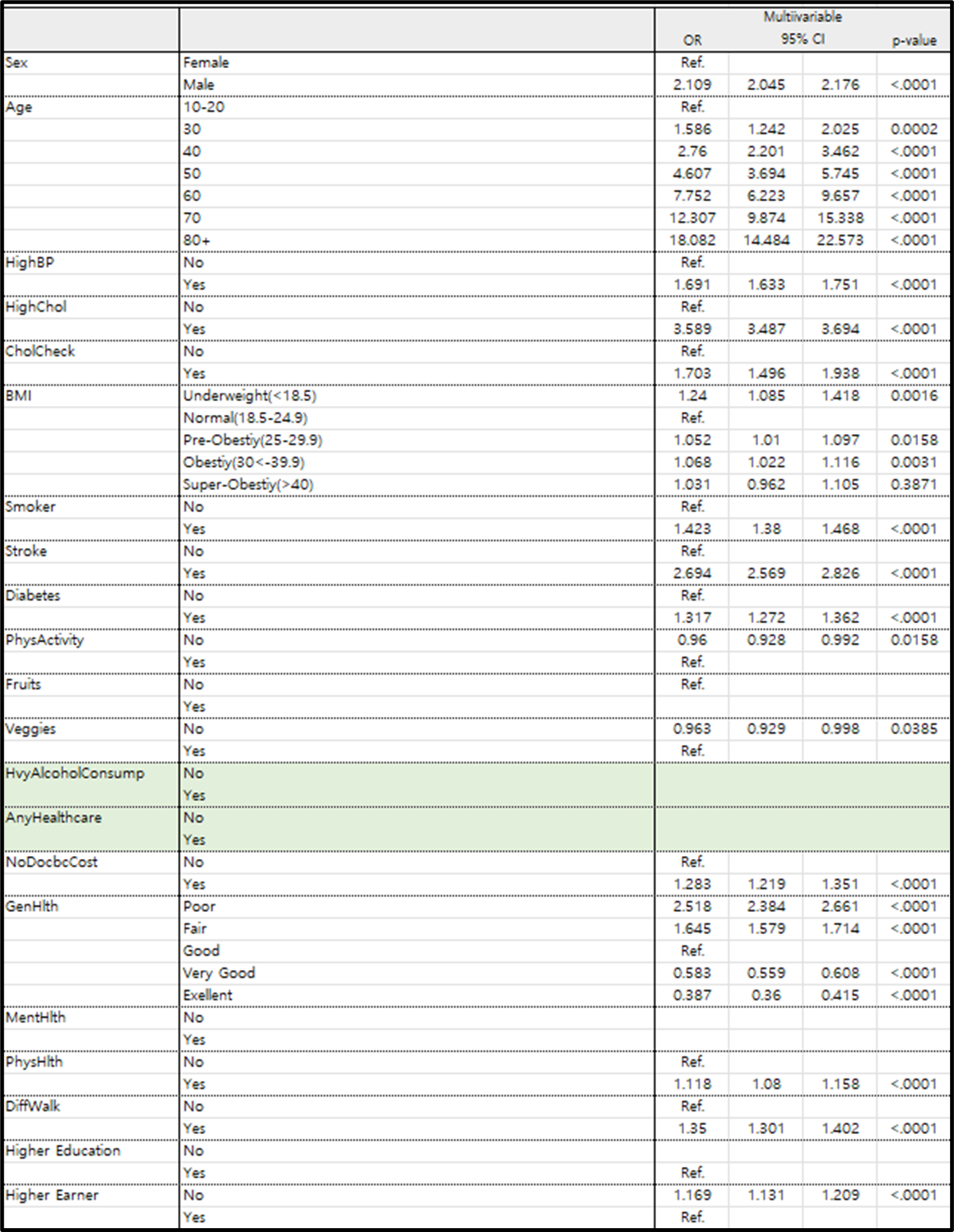
it seemed that recognizing one's own health status had a negative relationship with several other negative variables, and the age variable had a positive relationship with general vascular problems. Physical problems have been shown to be related to walking difficulties.

Simple visualization gives an intuitive concept, but it is not specific to support the basis of the argument.

Therefore, In order to capture a more accurate relationship, I would like to explain an analysis method that can confirm the influence of variables.

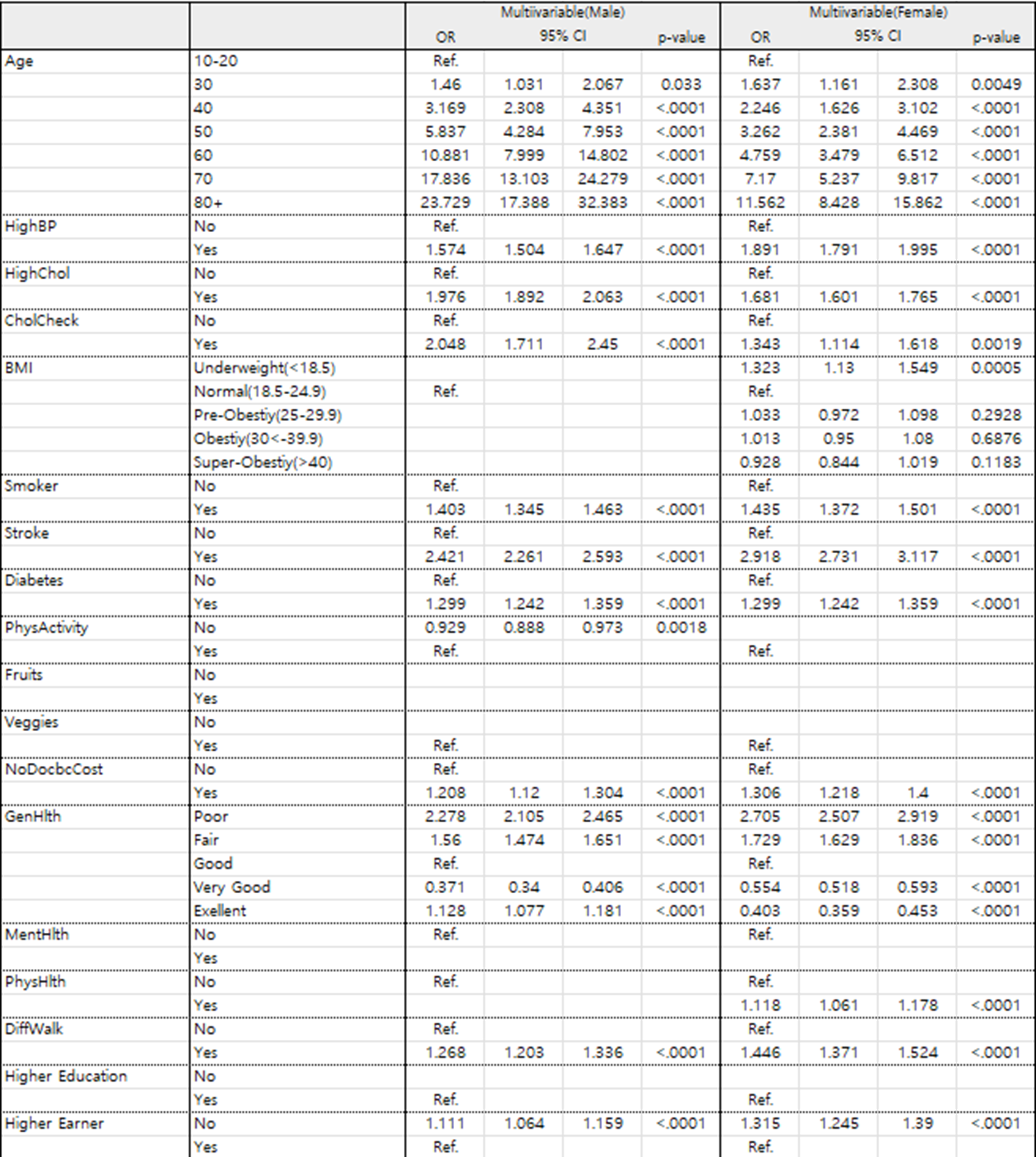


First, logistic regression analysis was performed first. Before constructing a polynomial regression model using variables, univariable logistic regression was conducted to confirm the influence of each variable, and the model construction was decided to proceed except for the two variables with low influence in this process. (The remaining variables seemed to have a significant influence on predicting heart disease to some extent.)



The polynomial logistic regression analysis using the entire variable is as above.

In the reduced model obtained through the variable selection method, it was confirmed that the Fruit, Mental Health, and High Education variables were not included, and the Physical Activity and veggies variables showed no significant influence.

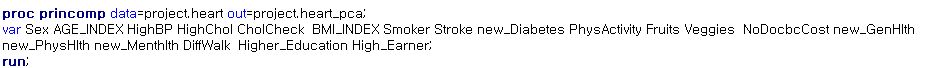


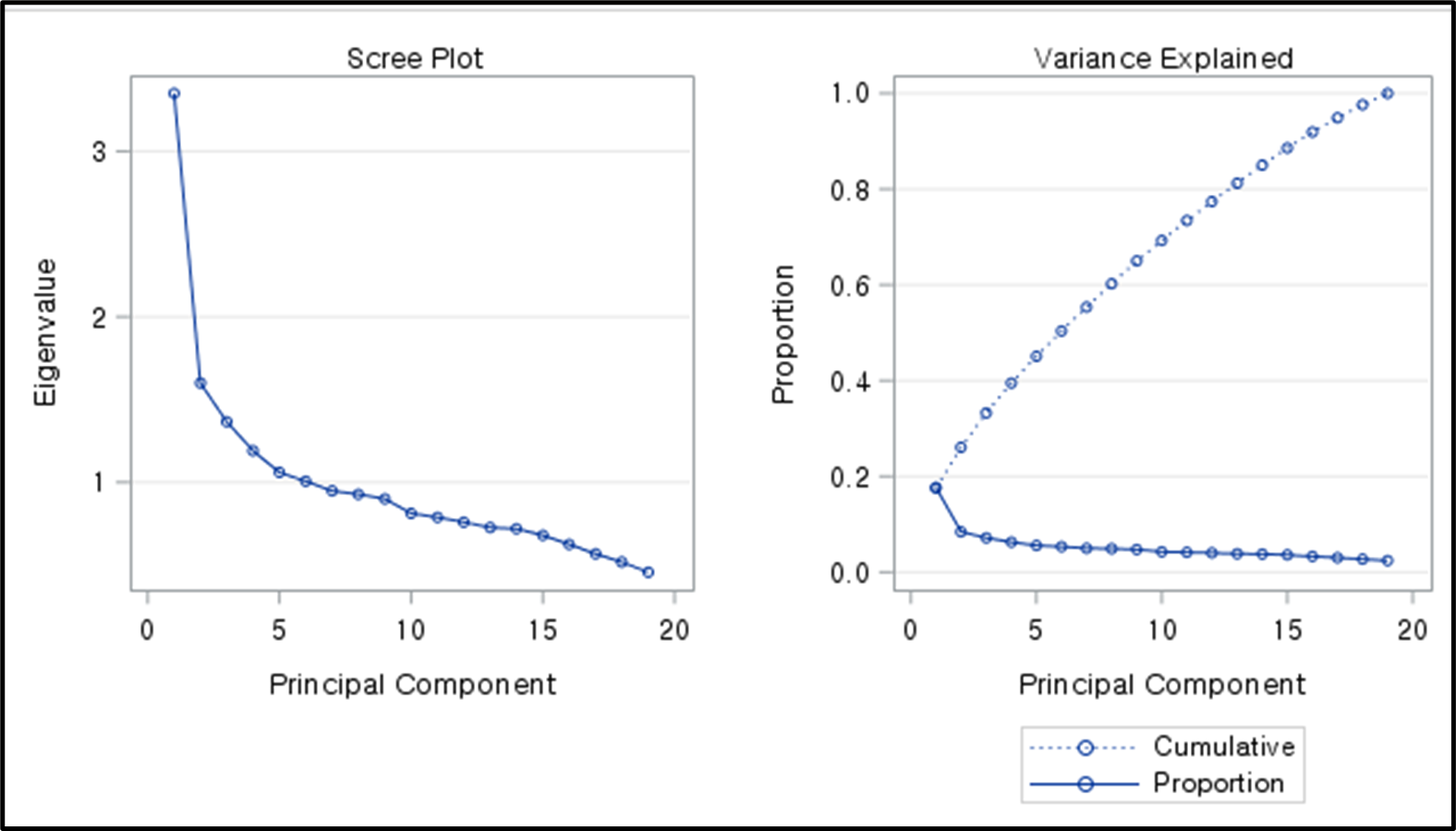
The table above is a polynomial logistic regression table according to gender.

Similar to the previous reduction model, both male and female women did not include Fruit, High Education, and Mental Health in the final model, and furthermore, veggies were not included in the final model, so this common part is considered not helpful in predicting regardless of gender.

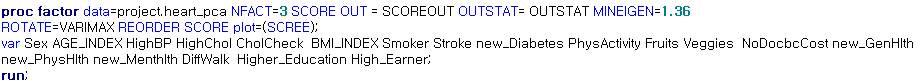
The next thing to proceed with is factor analysis. Through factor analysis, it was decided to compress the number of variables to derive only a few explanable parts.

In general, factor analysis focuses on how to calculate variance, and the method applied to this data is to apply PCA and rotate the axis around the covariance calculated through VARIMAX to find the insight for the new section.





In general, a criterion for configuring a factor is limited through a boundary value through a degree of a bent part of the eigen value through a screen plot, and when the eigen value is obtained by setting the lower limit of 1.36, three factors could be obtained and the calculated value is rotated on the axis.



Pattern values for the obtained factors and variables are as follows. Here, the variables to be included in the factor must be selected through cut-off.

I set the cutoff criterion at 0.4, and accordingly, the variables selected from each factor are as follows.

In the case of factor1, Physical Health Problem, Differences walking, Mental Health Problem, No meet doctor cause cost, factor2 is Age, High BP, High cholesterol. Diabetes, and finally, factors3 and factors3.



Through these variables selected for each factor, each factor was named.

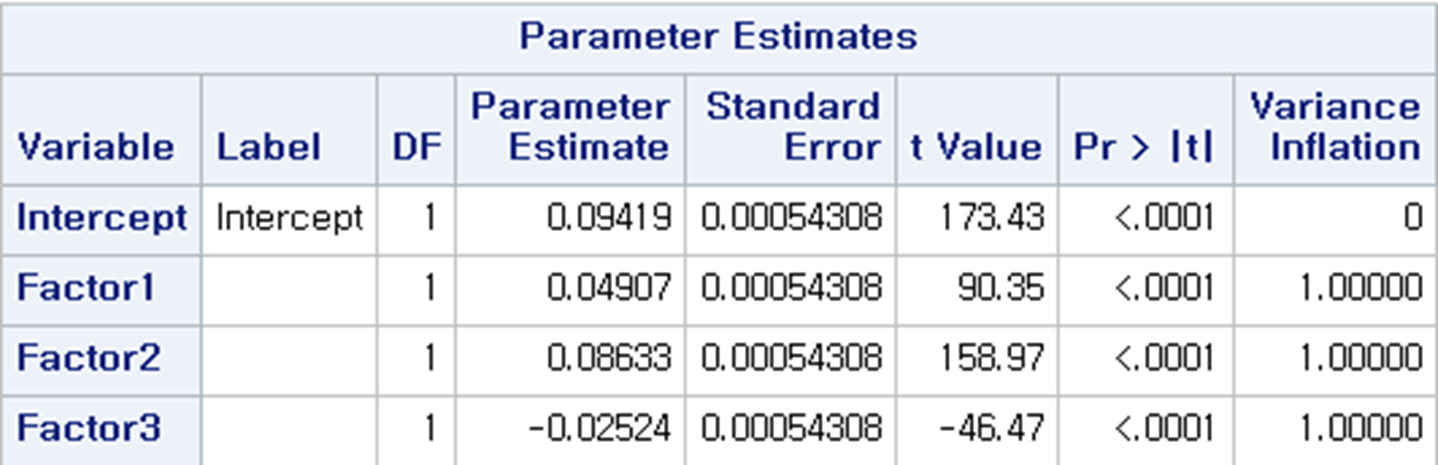
First of all, Factor 1 was called Low SES, but when looking at the variables in only one positive part, it may be said that it was judged only by looking at the noDocbcCost, but when looking at the variables that affect negatively, it seemed to be about various mental problems caused by low SES.

Second, Factor 2 is said to be in good health because it includes various physical problems.

Third, in the case of Factor 3, it was designated as food because there are veggies and fruits.



The following can be obtained when regression analysis is performed based on the selected factors.



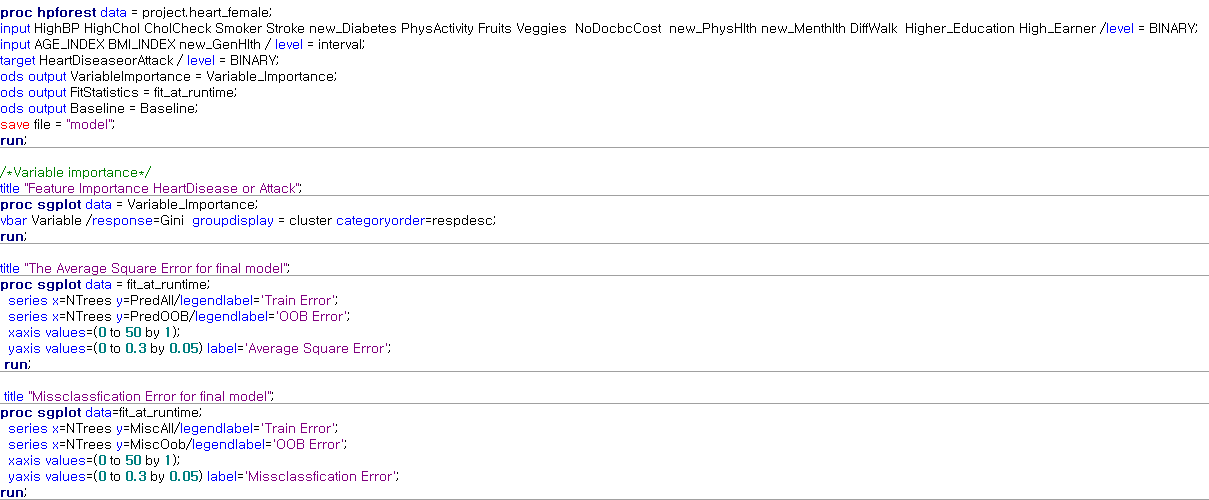
The interpretations obtained through the coefficients for the above are as follows.

When Factor 1 increases by 1, Heart Disease and Attack increases by 0.04907. In other words, it can be seen that the more physical and mental problems or difficulties that can occur due to weak socioeconomic aspects, the more heart disease and attack they are.

When Factor 2 increases by 1, Heart Disease and Attack increases by 0.08633. In other words, it can be seen that the more physical problems, especially aging and vascular problems, the more heart disease and attack they have.

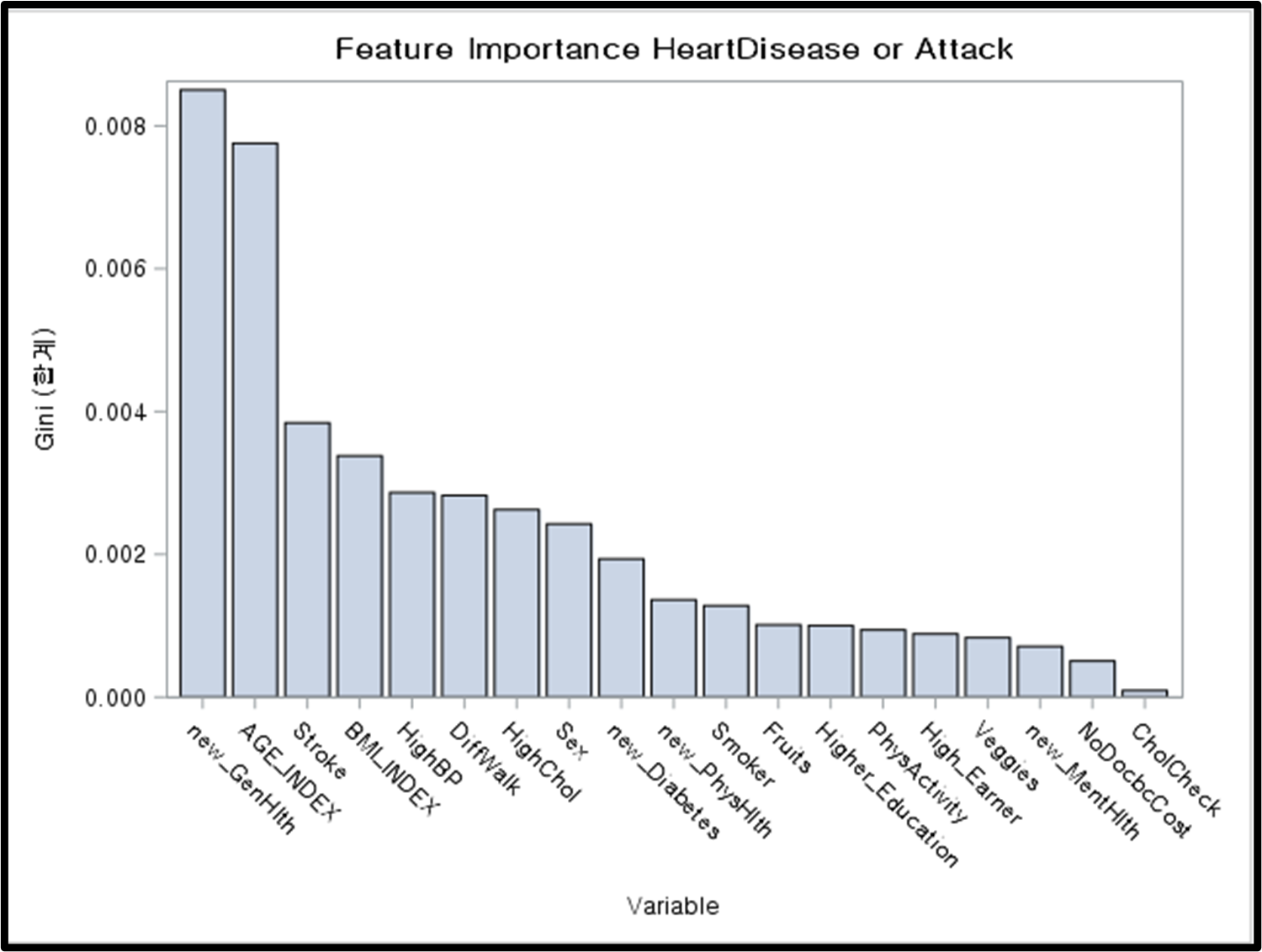
As Factor3 increases by 1, Heart Disease and Attack decreases by 0.02524. In other words, it can be seen that the more vegetables or fruits you eat, the less Heart Disease and Attack you eat.

The last analysis method is Random forest. Random Forest is a technique that facilitates decision-making and is used because it is suitable for classification problems. When these methods were used, these methods were chosen because the main variables in predicting could be identified.



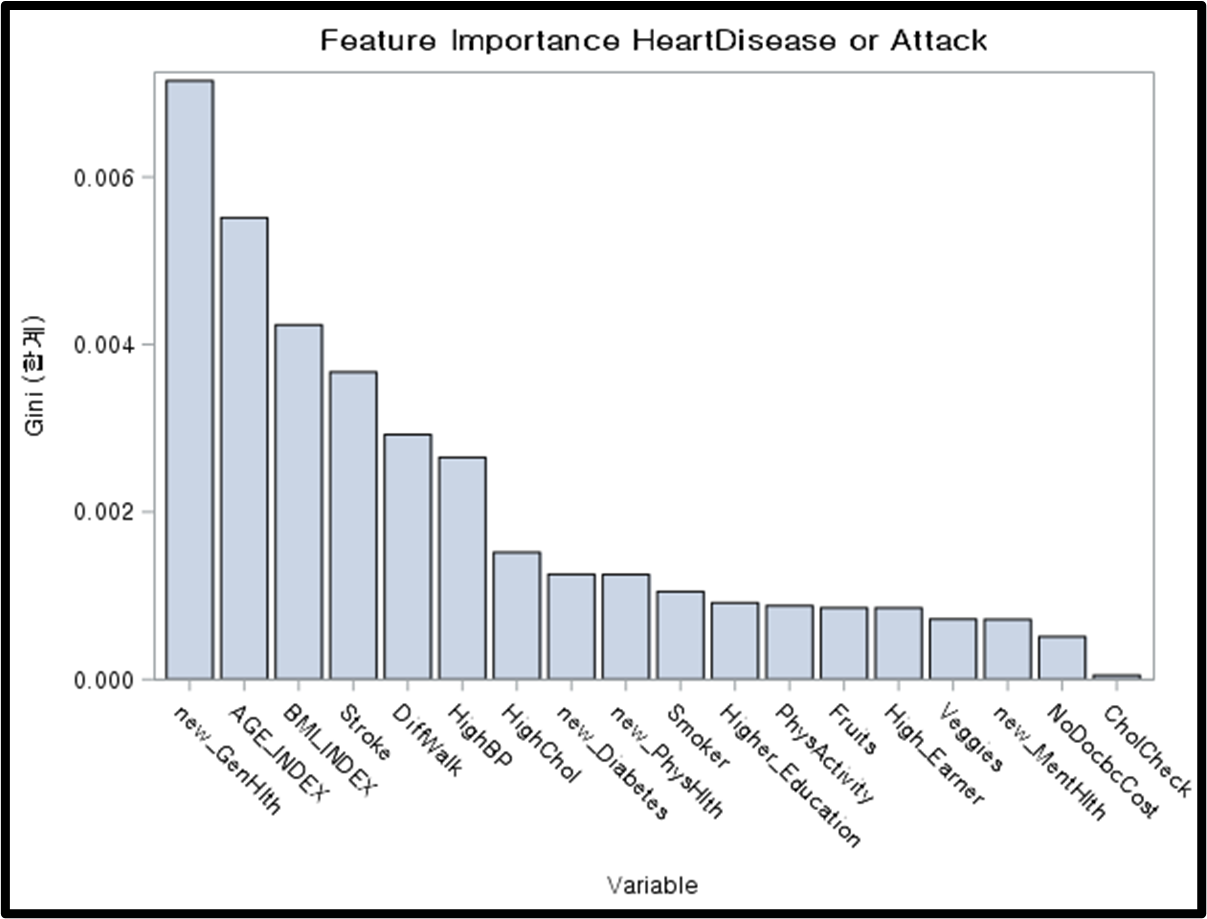
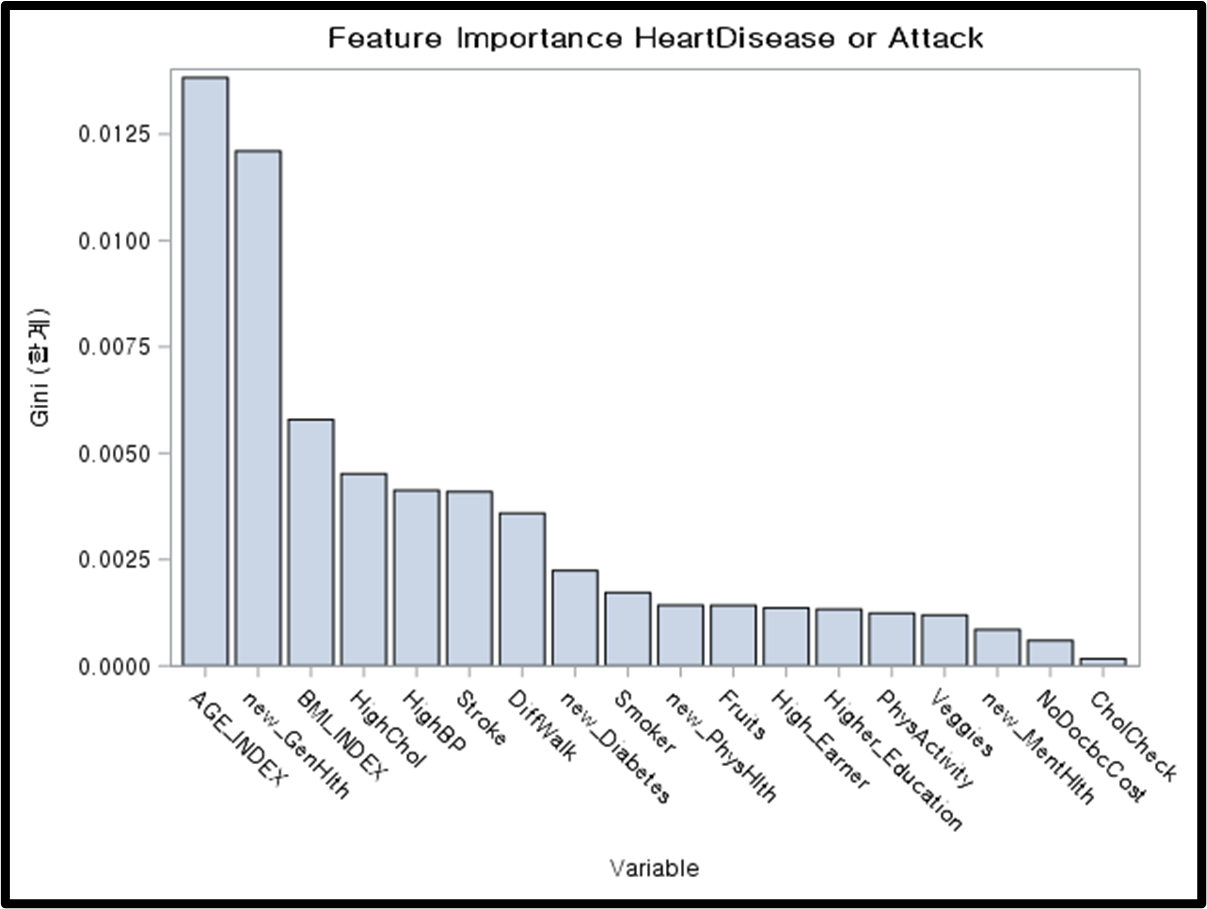
The degree of prediction in the default state was checked without adjusting other parameters.

When the entire variable is used, the estimated importance of each variable is as follows.



General Health, or perception of one's own health, was selected as the variable that had the most influence, and secondly, age seemed to have the most influence. In addition, the same variables such as stroke and BMI also influence, but they did not seem to show as much influence as the previous two variables.

Next, in order to find out how the importance of each variable changes when gender is divided, the model fit according to gender was confirmed.



It shows the importance of variables when the model is fitted through data on men on the left and women on the right. In the case of men, unlike the previous model, the Age variable was found to have the greatest influence, and secondly, it was confirmed that the General Health variable had the greatest influence. However, as in the previous model, these two variables showed the greatest influence in constructing the predictive diagram of the model. In the case of women, General Health was selected as the most important variable as in the previous model, and the AGE variable was selected next. However, the proportion of the two variables was different, and when looking at the remaining variables other than the two variables, the proportion of importance seemed to be evenly distributed. In other words, it was confirmed that certain variables did not overwhelmingly influence the predictability of the model.

# 4. Conclusion & Discussion

1. In predicting Heart Disease, the importance of variables such as Age and General Health (Awareness of Health) was largely confirmed, and in addition, it was confirmed that there was a physical problem with variables representing SES.

2. However, in common, it did not matter whether or not the Mental Health problem, High Education, and Heavy Alcohol consumption, received benefit for Health insurance.

**Task 1**> It may differ depending on which model will be mainly used. Rather than building a model with simple numbers as it is, I think building a binary structure with a yes or no or a continuous value can increase data accuracy and interpretation in building a model.

**Task 2**> Yes, Parts such as basic demographic variables such as age and gender and General health, which describe self-health awareness, can be judged as major variables in predicting heart disease, physical problems (stroke, HighBP, HighCol, Diabetes) can be considered for careful judgment, and SES can be considered

### **Appendix**

1. Preprocessing(Python)

| **## 1. Get the data**  **#imports**  **import os**  **import pandas as pd**  **import random**  **random.seed(1)**  **#read in the dataset (select 2015)**  **year = '2015'**  **brfss\_2015\_dataset = pd.read\_csv(f'../input/behavioral-risk-factor-surveillance-system/{year}.csv')**  **#How many rows and columns**  **brfss\_2015\_dataset.shape**  **#check that the data loaded in is in the correct format**  **pd.set\_option('display.max\_columns', 500)**  **brfss\_2015\_dataset.head()**  **# select specific columns**  **brfss\_df\_selected = brfss\_2015\_dataset[['\_MICHD',**  **'\_RFHYPE5',**  **'TOLDHI2', '\_CHOLCHK',**  **'\_BMI5',**  **'SMOKE100',**  **'CVDSTRK3', 'DIABETE3',**  **'\_TOTINDA',**  **'\_FRTLT1', '\_VEGLT1',**  **'\_RFDRHV5',**  **'HLTHPLN1', 'MEDCOST',**  **'GENHLTH', 'MENTHLTH', 'PHYSHLTH', 'DIFFWALK',**  **'SEX', '\_AGEG5YR', 'EDUCA', 'INCOME2' ]]**  **## 2. Clean the data**  **# 2.1 Drop missing values**  **#Drop Missing Values - knocks 100,000 rows out right away**  **brfss\_df\_selected = brfss\_df\_selected.dropna()**  **brfss\_df\_selected.shape**  **#2.2 Modify and clean the values to be more suitable to ML Algorithms**  **# \_MICHD**  **#Change 2 to 0 because this means did not have MI or CHD**  **brfss\_df\_selected['\_MICHD'] = brfss\_df\_selected['\_MICHD'].replace({2: 0})**  **brfss\_df\_selected.\_MICHD.unique()**  **#1 \_RFHYPE5**  **#Change 1 to 0 so it represetnts No high blood pressure and 2 to 1 so it represents high blood pressure**  **brfss\_df\_selected['\_RFHYPE5'] = brfss\_df\_selected['\_RFHYPE5'].replace({1:0, 2:1})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.\_RFHYPE5 != 9]**  **brfss\_df\_selected.\_RFHYPE5.unique()**  **#2 TOLDHI2**  **# Change 2 to 0 because it is No**  **# Remove all 7 (dont knows)**  **# Remove all 9 (refused)**  **brfss\_df\_selected['TOLDHI2'] = brfss\_df\_selected['TOLDHI2'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.TOLDHI2 != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.TOLDHI2 != 9]**  **brfss\_df\_selected.TOLDHI2.unique()**  **#3 \_CHOLCHK**  **# Change 3 to 0 and 2 to 0 for Not checked cholesterol in past 5 years**  **# Remove 9**  **brfss\_df\_selected['\_CHOLCHK'] = brfss\_df\_selected['\_CHOLCHK'].replace({3:0,2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.\_CHOLCHK != 9]**  **brfss\_df\_selected.\_CHOLCHK.unique()**  **#4 \_BMI5 (no changes, just note that these are BMI \* 100. So for example a BMI of 4018 is really 40.18)**  **brfss\_df\_selected['\_BMI5'] = brfss\_df\_selected['\_BMI5'].div(100).round(0)**  **brfss\_df\_selected.\_BMI5.unique()**  **#5 SMOKE100**  **# Change 2 to 0 because it is No**  **# Remove all 7 (dont knows)**  **# Remove all 9 (refused)**  **brfss\_df\_selected['SMOKE100'] = brfss\_df\_selected['SMOKE100'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.SMOKE100 != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.SMOKE100 != 9]**  **brfss\_df\_selected.SMOKE100.unique()**  **#6 CVDSTRK3**  **# Change 2 to 0 because it is No**  **# Remove all 7 (dont knows)**  **# Remove all 9 (refused)**  **brfss\_df\_selected['CVDSTRK3'] = brfss\_df\_selected['CVDSTRK3'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.CVDSTRK3 != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.CVDSTRK3 != 9]**  **brfss\_df\_selected.CVDSTRK3.unique()**  **#7 DIABETE3**  **# going to make this ordinal. 0 is for no diabetes or only during pregnancy, 1 is for pre-diabetes or borderline diabetes, 2 is for yes diabetes**  **# Remove all 7 (dont knows)**  **# Remove all 9 (refused)**  **brfss\_df\_selected['DIABETE3'] = brfss\_df\_selected['DIABETE3'].replace({2:0, 3:0, 1:2, 4:1})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.DIABETE3 != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.DIABETE3 != 9]**  **brfss\_df\_selected.DIABETE3.unique()**  **#8 \_TOTINDA**  **# 1 for physical activity**  **# change 2 to 0 for no physical activity**  **# Remove all 9 (don't know/refused)**  **brfss\_df\_selected['\_TOTINDA'] = brfss\_df\_selected['\_TOTINDA'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.\_TOTINDA != 9]**  **brfss\_df\_selected.\_TOTINDA.unique()**  **#9 \_FRTLT1**  **# Change 2 to 0. this means no fruit consumed per day. 1 will mean consumed 1 or more pieces of fruit per day**  **# remove all dont knows and missing 9**  **brfss\_df\_selected['\_FRTLT1'] = brfss\_df\_selected['\_FRTLT1'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.\_FRTLT1 != 9]**  **brfss\_df\_selected.\_FRTLT1.unique()**  **#10 \_VEGLT1**  **# Change 2 to 0. this means no vegetables consumed per day. 1 will mean consumed 1 or more pieces of vegetable per day**  **# remove all dont knows and missing 9**  **brfss\_df\_selected['\_VEGLT1'] = brfss\_df\_selected['\_VEGLT1'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.\_VEGLT1 != 9]**  **brfss\_df\_selected.\_VEGLT1.unique()**  **#11 \_RFDRHV5**  **# Change 1 to 0 (1 was no for heavy drinking). change all 2 to 1 (2 was yes for heavy drinking)**  **# remove all dont knows and missing 9**  **brfss\_df\_selected['\_RFDRHV5'] = brfss\_df\_selected['\_RFDRHV5'].replace({1:0, 2:1})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.\_RFDRHV5 != 9]**  **brfss\_df\_selected.\_RFDRHV5.unique()**  **#12 HLTHPLN1**  **# 1 is yes, change 2 to 0 because it is No health care access**  **# remove 7 and 9 for don't know or refused**  **brfss\_df\_selected['HLTHPLN1'] = brfss\_df\_selected['HLTHPLN1'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.HLTHPLN1 != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.HLTHPLN1 != 9]**  **brfss\_df\_selected.HLTHPLN1.unique()**  **#13 MEDCOST**  **# Change 2 to 0 for no, 1 is already yes**  **# remove 7 for don/t know and 9 for refused**  **brfss\_df\_selected['MEDCOST'] = brfss\_df\_selected['MEDCOST'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.MEDCOST != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.MEDCOST != 9]**  **brfss\_df\_selected.MEDCOST.unique()**  **#14 GENHLTH**  **# This is an ordinal variable that I want to keep (1 is Excellent -> 5 is Poor)**  **# Remove 7 and 9 for don't know and refused**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.GENHLTH != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.GENHLTH != 9]**  **brfss\_df\_selected.GENHLTH.unique()**  **#15 MENTHLTH**  **# already in days so keep that, scale will be 0-30**  **# change 88 to 0 because it means none (no bad mental health days)**  **# remove 77 and 99 for don't know not sure and refused**  **brfss\_df\_selected['MENTHLTH'] = brfss\_df\_selected['MENTHLTH'].replace({88:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.MENTHLTH != 77]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.MENTHLTH != 99]**  **brfss\_df\_selected.MENTHLTH.unique()**  **#16 PHYSHLTH**  **# already in days so keep that, scale will be 0-30**  **# change 88 to 0 because it means none (no bad mental health days)**  **# remove 77 and 99 for don't know not sure and refused**  **brfss\_df\_selected['PHYSHLTH'] = brfss\_df\_selected['PHYSHLTH'].replace({88:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.PHYSHLTH != 77]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.PHYSHLTH != 99]**  **brfss\_df\_selected.PHYSHLTH.unique()**  **#17 DIFFWALK**  **# change 2 to 0 for no. 1 is already yes**  **# remove 7 and 9 for don't know not sure and refused**  **brfss\_df\_selected['DIFFWALK'] = brfss\_df\_selected['DIFFWALK'].replace({2:0})**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.DIFFWALK != 7]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.DIFFWALK != 9]**  **brfss\_df\_selected.DIFFWALK.unique()**  **#18 SEX**  **# in other words - is respondent male (somewhat arbitrarily chose this change because men are at higher risk for heart disease)**  **# change 2 to 0 (female as 0). Male is 1**  **brfss\_df\_selected['SEX'] = brfss\_df\_selected['SEX'].replace({2:0})**  **brfss\_df\_selected.SEX.unique()**  **#19 \_AGEG5YR**  **# already ordinal. 1 is 18-24 all the way up to 13 wis 80 and older. 5 year increments.**  **# remove 14 because it is don't know or missing**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.\_AGEG5YR != 14]**  **brfss\_df\_selected.\_AGEG5YR.unique()**  **#20 EDUCA**  **# This is already an ordinal variable with 1 being never attended school or kindergarten only up to 6 being college 4 years or more**  **# Scale here is 1-6**  **# Remove 9 for refused:**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.EDUCA != 9]**  **brfss\_df\_selected.EDUCA.unique()**  **#21 INCOME2**  **# Variable is already ordinal with 1 being less than $10,000 all the way up to 8 being $75,000 or more**  **# Remove 77 and 99 for don't know and refused**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.INCOME2 != 77]**  **brfss\_df\_selected = brfss\_df\_selected[brfss\_df\_selected.INCOME2 != 99]**  **brfss\_df\_selected.INCOME2.unique()**  **#Check the shape of the dataset now: We have 253,680 cleaned rows and 22 columns (1 of which is our dependent variable)**  **brfss\_df\_selected.shape**  **#Check Class Sizes of the heart disease column**  **brfss\_df\_selected.groupby(['\_MICHD']).size()**  **# 3. Make feature names more readable**  **#Rename the columns to make them more readable**  **brfss = brfss\_df\_selected.rename(columns =**  **{'\_MICHD':'HeartDiseaseorAttack','\_RFHYPE5':'HighBP', 'TOLDHI2':'HighChol',**  **'\_CHOLCHK':'CholCheck','\_BMI5':'BMI','SMOKE100':'Smoker','CVDSTRK3':'Stroke',**  **'DIABETE3':'Diabetes','\_TOTINDA':'PhysActivity','\_FRTLT1':'Fruits',**  **'\_VEGLT1':"Veggies",'\_RFDRHV5':'HvyAlcoholConsump','HLTHPLN1':'AnyHealthcare',**  **'MEDCOST':'NoDocbcCost','GENHLTH':'GenHlth', 'MENTHLTH':'MentHlth', 'PHYSHLTH':'PhysHlth', 'DIFFWALK':'DiffWalk','SEX':'Sex', '\_AGEG5YR':'Age', 'EDUCA':'Education', 'INCOME2':'Income' })**  **#Check how many respondents have had heart disease or a heart attack. Note the class imbalance!**  **brfss.groupby(['HeartDiseaseorAttack']).size()**  **#4. Save to csv**  **#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***  **brfss.to\_csv('heart\_disease\_health\_indicators\_BRFSS2015.csv', sep=",", index=False)**  **#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*** |
| --- |

2. Data Manipulation & Analysis

| *libname project "C:\Users\jun\Dropbox\graduate\_school\2021\_fall\_semetser\Statistical Computing\term\_project";  /\*Load data\*/  proc import datafile = "C:\Users\jun\Dropbox\graduate\_school\2021\_fall\_semetser\Statistical Computing\term\_project\heart\_disease\_health\_indicators\_BRFSS2015.csv"  /\*proc import datafile = "/home/u42028077/final\_term/heart\_disease\_health\_indicators\_BRFSS2015.csv"\*/ dbms = csv out = project.heart replace; getnames=Yes; run;  /\* Checking the contents of the dataset \*/ proc contents data = project.heart varnum; run;  data project.heart; set project.heart; label   HeartDiseaseorAttack = "0=No,1=Yes"  Age = "1=18-24,2=25-29,3=30-34,4=35-39,5=40-44,6=45-49,7=50-54,8=55-59,9=60-64,10=65-69,11=70-74,12=75-79,13=80+"  SEX ="0=Female, 1=Male"  AGE\_INDEX = "1=<30,2=30's,3=40's,4=50's,5=60's,6=70's,7=80+"  HighBP = "0=No,1=Yes"  HighChol = "0=No,1=Yes"  CholCheck = "0=No,1=Yes"  Smoker= "0=No,1=Yes"  Stroke= "0=No,1=Yes"  BMI = "Body Mass Index"  Diabetes="0=No,1='Pre-Diabetes,2 = Diabetes"  PhysActivity= "0=No,1=Yes"  Fruits= "0=No,1=Yes"  Veggies = "0=No,1=Yes"  HvyAlcoholConsump= "0=No,1=Yes"  AnyHealthcare="0=No,1=Yes"  NoDocbcCost= "0=No,1=Yes"  GenHlth= "1=Excellent,2=Very Good, 3=Good, 4=Fair, 5=Poor"  MentHlth = "how many days during the past 30 days was your physical health not good"  PhysHlth = "how many days during the past 30 days was your mental health not good"  DiffWalk= "0=No,1=Yes"  Education = "1=Less than Kindergarten, 2=Elementary. 3=Some high school, 4=High School Grauated, 5 = College or Tehnical school, 6 = above college graduage"  Income = " 1=<$10,000,2=<$15,000,3=<$20,000, 4=<$25,000, 5=<$35,000,6=<$50,000,7=<$75,000,8=>$75,000, " ; run;*  */\*예측 변수 정렬\*/*  *proc sort data = project.heart; by Descending HeartDiseaseorAttack ; run;*  *proc print data = project.heart(obs=50); run;*  */\*Summary Statistics\*/*  *proc means data=project.heart mean median mode std var min max;*  *run;*  */\*Finding Missing Values\*/*  *proc means data=project.heart nmiss;*  *run;*  */\*-------BMI 수치가 이상해서 확인 ------\*/*  */\*거식증인 환자의 경우나 혹은 초고도비만인 사람이 존재할 가능성이 충분히 존재하기 때문에 데이터의 신빙성의 문제라고 보기에는 어려움\*/*  */\*253680\*/*  *proc univariate data = project.heart plot;*  *class SEX;*  *var BMI;*  *run;*  *proc sgpanel data = project.heart;*  *title 'BMI faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram BMI / group= HeartDiseaseorAttack;*  *run;*  */\*-------General Health awareness by self ------\*/*  *proc sgpanel data = project.heart;*  *title 'Health Awareness by self faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram GenHlth / group=HeartDiseaseorAttack;*  *run;*  */\*-------Menthal Health & Physicsl Health ------\*/*  *proc sgpanel data = project.heart;*  *title 'Mental Health Problem faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram MentHlth / group=HeartDiseaseorAttack;*  *run;*  *proc sgpanel data = project.heart;*  *title 'Physcial Health Problem faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram PhysHlth / group=HeartDiseaseorAttack;*  *run;*  */\*-------Diabetes ------\*/*  *proc sgpanel data = project.heart;*  *title 'Diabetes faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram Diabetes / group=HeartDiseaseorAttack;*  *run;*  */\*-------Education ------\*/*  *proc sgpanel data = project.heart;*  *title 'Education faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram Education / group=HeartDiseaseorAttack;*  *run;*  */\*-------Income ------\*/*  *proc sgpanel data = project.heart;*  *title 'Income faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram Income / group=HeartDiseaseorAttack;*  *run;*  *proc sgpanel data = project.heart;*  *title 'Age faceted by Heart Disease or Attack & Sex';*  *panelby Sex;*  *histogram Age / group=HeartDiseaseorAttack;*  *run;*  */\*미국 기준 BMI수정 AGE index 수정(5세 단위에서 10세단위로 )\*/*  *data project.heart(drop = BMI AGE Diabetes GenHlth education income MentHlth PhysHlth) ; set project.heart ;*  */\*1- Underweight 2-Normal 3- Pre Obesity 4-Obestiy 5 Super-Obestiy\*/*  *if BMI <18.5 then BMI\_INDEX = 1;*  *else if BMI >= 18.5 and BMI <25 then BMI\_INDEX =2;*  *else if BMI >= 25 and BMI <30 then BMI\_INDEX = 3;*  *else if BMI >= 30 and BMI <40 then BMI\_INDEX = 4;*  *else BMI\_INDEX = 5;*  */\*1-10&20's(18<) 2 - 30's 3-40's 4-50's 5-60's 6-70's 7-80+\*/*  *If AGE =1 or AGE =2 then AGE\_INDEX =1;*  *else if AGE =3 or AGE =4 then AGE\_INDEX =2;*  *else if AGE =5 or AGE =6 then AGE\_INDEX =3;*  *else if AGE =7 or AGE =8 then AGE\_INDEX =4;*  *else if AGE =9 or AGE =10 then AGE\_INDEX =5;*  *else if AGE =11 or AGE =12 then AGE\_INDEX =6;*  *else AGE\_INDEX =7;*  */\*Pre-Diabetes 의 90%가 Diabetes으로 발전하는 것을 고려하여 하나로 통합\*/*  *if Diabetes = 0 then new\_Diabetes = 0;*  *else new\_Diabetes =1;*    */\*5- Excellent 4-very good 3-good 2-fair 1-poor\*/*  *if GenHlth = 1 then new\_GenHlth = 5;*  *else if GenHlth = 2 then new\_GenHlth = 4;*  *else if GenHlth = 3 then new\_GenHlth = 3;*  *else if GenHlth = 4 then new\_GenHlth = 2;*  *else new\_GenHlth = 1;*  */\* 실질적인 의무교육은 중학교 까지이지만 미국의 고등학교 졸업율이 70%감안하였을 때 의무교육으로 간주함 (대학교는 완전한 선택사항이므로) \*/*  *if education in (1,2,3,4) then Higher\_Education =0;*  *else Higher\_Education =1;*  *if income in (1,2,3,4,5,6) then High\_Earner = 0;*  *else High\_Earner =1;*  */\* min-max scaling 또는 standardized 해도 되는데 쉬운 비교를 위해서 factotr 변수로 변환\*/*  *if MentHlth in (0,1,2,3,4,5) then new\_MentHlth = 0;*  *else new\_MentHlth = 1;*  *if PhysHlth in (0,1,2,3,4,5) then new\_PhysHlth = 0;*  *else new\_PhysHlth = 1;*  *run;*  */\*To check Freqeuncy \*/*  *proc freq data = project.heart;*  *tables*  *Sex\* HeartDiseaseorAttack*  *AGE\_INDEX \* HeartDiseaseorAttack*  *HighBP \* HeartDiseaseorAttack*  *HighChol \* HeartDiseaseorAttack*  *CholCheck \* HeartDiseaseorAttack*  *BMI\_INDEX \* HeartDiseaseorAttack*  *Smoker \* HeartDiseaseorAttack*  *Stroke \* HeartDiseaseorAttack*  *new\_Diabetes \* HeartDiseaseorAttack*  *PhysActivity \* HeartDiseaseorAttack*  *Fruits \* HeartDiseaseorAttack*  *Veggies \* HeartDiseaseorAttack*  *HvyAlcoholConsump \* HeartDiseaseorAttack*  *AnyHealthcare \* HeartDiseaseorAttack*  *NoDocbcCost \* HeartDiseaseorAttack*  *new\_GenHlth \* HeartDiseaseorAttack*  *new\_MentHlth\* HeartDiseaseorAttack*  *new\_PhysHlth\* HeartDiseaseorAttack*  *DiffWalk \* HeartDiseaseorAttack*  *Higher\_Education \* HeartDiseaseorAttack*  *High\_Earner \* HeartDiseaseorAttack*  */ nocol norow nocum nopercent*  *;*  *run;*  */\*bar graph \*/*  *%macro check\_hist(out=,title=,);*  *proc sgplot data = project.heart;*  *vbar &out. / group= HeartDiseaseorAttack groupdisplay = cluster;*  *title &title.;*  *run;*  *%mend;*  *%check\_hist(out=SEX,title="Sex VS Heart Disease or Attack");*  *%check\_hist(out=AGE\_INDEX,title="AGE\_INDEX VS Heart Disease or Attack");*  *%check\_hist(out=HighBP,title="HighBP VS Heart Disease or Attack");*  *%check\_hist(out=HighChol,title="HighChol VS Heart Disease or Attack");*  *%check\_hist(out=CholCheck,title="CholCheck VS Heart Disease or Attack");*  *%check\_hist(out=BMI\_INDEX,title="BMI\_INDEX VS Heart Disease or Attack");*  *%check\_hist(out=Smoker,title="Smoker VS Heart Disease or Attack");*  *%check\_hist(out=Stroke,title="Stroke VS Heart Disease or Attack");*  *%check\_hist(out=new\_Diabetes,title="new\_Diabetes VS Heart Disease or Attack");*  *%check\_hist(out=PhysActivity,title="PhysActivity VS Heart Disease or Attack");*  *%check\_hist(out=Fruits,title="Fruits VS Heart Disease or Attack");*  *%check\_hist(out=Veggies,title="Veggies VS Heart Disease or Attack");*  *%check\_hist(out=HvyAlcoholConsump,title="HvyAlcoholConsump VS Heart Disease or Attack");*  *%check\_hist(out=AnyHealthcare,title="AnyHealthcare VS Heart Disease or Attack");*  *%check\_hist(out=NoDocbcCost,title="NoDocbcCost VS Heart Disease or Attack");*  *%check\_hist(out=new\_GenHlth,title="GenHlth VS Heart Disease or Attack");*  *%check\_hist(out=new\_MentHlth,title="MentHlth VS Heart Disease or Attack");*  *%check\_hist(out=new\_PhysHlth,title="PhysHlth VS Heart Disease or Attack");*  *%check\_hist(out=Higher\_Education,title="Higher\_Education VS Heart Disease or Attack");*  *%check\_hist(out=Higher\_Earner,title="Higher\_Earner VS Heart Disease or Attack");*  */\*mosaic graph \*/*  *%macro check\_mosaic(out=,title=,);*  *ods graphics on;*  *proc freq data=project.heart;*  *tables HeartDiseaseorAttack\*&out. / norow nofreq plots=MOSAIC;*  *title &title.;*  *run;*  *%mend;*  *%check\_mosaic(out=SEX,title="Distribution of Heart Disease or Attack by SEX");*  *%check\_mosaic(out=HighBP,title="Distribution of Heart Disease or Attack by HighBP");*  *%check\_mosaic(out=HighChol,title="Distribution of Heart Disease or Attack by HighChol");*  *%check\_mosaic(out=CholCheck,title="Distribution of Heart Disease or Attack by CholCheck");*  *%check\_mosaic(out=Smoker,title="Distribution of Heart Disease or Attack by Smoker");*  *%check\_mosaic(out=Stroke,title="Distribution of Heart Disease or Attack by Stroke");*  *%check\_mosaic(out=new\_Diabetes,title="Distribution of Heart Disease or Attack by Diabetes");*  *%check\_mosaic(out=PhysActivity,title="Distribution of Heart Disease or Attack by PhysActivity");*  *%check\_mosaic(out=Fruits,title="Distribution of Heart Disease or Attack by Fruits");*  *%check\_mosaic(out=Veggies,title="Distribution of Heart Disease or Attack by Veggies");*  *%check\_mosaic(out=HvyAlcoholConsump,title="Distribution of Heart Disease or Attack by HvyAlcoholConsump");*  *%check\_mosaic(out=NoDocbcCost,title="Distribution of Heart Disease or Attack by NoDocbcCost");*  *%check\_mosaic(out=new\_MentHlth,title="Distribution of Heart Disease or Attack by MentHlth ");*  *%check\_mosaic(out=new\_PhysHlth,title="Distribution of Heart Disease or Attack by PhysHlth");*  *%check\_mosaic(out=Higher\_Education,title="Distribution of Heart Disease or Attack by Higher Education ");*  *%check\_mosaic(out=High\_Earner,title="Distribution of Heart Disease or Attack by High Earner");*  *%check\_mosaic(out=AGE\_INDEX,title="Distribution of Heart Disease or Attack by AGE\_INDEX");*  *%check\_mosaic(out=BMI\_INDEX,title="Distribution of Heart Disease or Attack by BMI\_INDEX");*  *%check\_mosaic(out=new\_GenHlth,title="Distribution of Heart Disease or Attack by GenHlth");*  */\*To check correlation \*/*  *proc template;*  *define statgraph corrHeatmap;*  *dynamic \_Title;*  *begingraph;*  *entrytitle \_Title;*  *rangeattrmap name='map';*  *range -1 - 1 / rangecolormodel=(cxD8B365 cxF5F5F5 cx5AB4AC);*  *endrangeattrmap;*  *rangeattrvar var=r attrvar=r attrmap='map';*  *layout overlay /*  *xaxisopts=(display=(line ticks tickvalues))*  *yaxisopts=(display=(line ticks tickvalues));*  *heatmapparm x = x y = y colorresponse = r / xbinaxis=false ybinaxis=false*  *colormodel=THREECOLORRAMP name = "heatmap" display=all;*  *continuouslegend "heatmap" /*  *orient = vertical location = outside title="Pearson Correlation";*  *endlayout;*  *endgraph;*  *end;*  *run;*  */\* Prepare the correlations coeff matrix: Pearson's r method \*/*  *%macro prepCorrData(in=,out=);*  */\* Run corr matrix for input data, all numeric vars \*/*  *proc corr data=&in. noprint*  *pearson*  *outp=work.\_tmpCorr*  *vardef=df;*  *run;*  */\* prep data for heatmap \*/*  *data &out.;*  *keep x y r;*  *set work.\_tmpCorr(where=(\_TYPE\_="CORR"));*  *array v{\*} \_numeric\_;*  *x = \_NAME\_;*  *do i = dim(v) to 1 by -1;*  *y = vname(v(i));*  *r = v(i);*  */\* creates a diagonally sparse matrix \*/*  *if (i<\_n\_) then*  *r=.;*  *output;*  *end;*  *run;*  *proc datasets lib=work nolist nowarn;*  *delete \_tmpcorr;*  *quit;*  *%mend;*  */\* Build the graphs \*/*  *%prepCorrData(in=project.heart,out=heart\_r);*  *proc sgrender data=heart\_r template=corrHeatmap;*  *dynamic \_title="Corr matrix for project.heart";*  *run;*  */\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/*  */\* Find Significant Risk Factor and Assessment \*/*  */\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/*  */\*------------------------Logistic--------------------------------\*/*  */\*univarate\*/*  *proc logistic data=project.heart descending; class Sex(ref='0');model HeartDiseaseorAttack = Sex/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class AGE\_INDEX(ref='1');model HeartDiseaseorAttack = AGE\_INDEX/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class HighBP(ref='0');model HeartDiseaseorAttack = HighBP/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class HighChol(ref='0');model HeartDiseaseorAttack = HighChol/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class CholCheck(ref='0');model HeartDiseaseorAttack = CholCheck/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class BMI\_INDEX(ref='2');model HeartDiseaseorAttack = BMI\_INDEX/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class Smoker(ref='0');model HeartDiseaseorAttack = Smoker/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class Stroke(ref='0');model HeartDiseaseorAttack = Stroke/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class new\_Diabetes(ref='0');model HeartDiseaseorAttack = new\_Diabetes/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class PhysActivity(ref='1');model HeartDiseaseorAttack = PhysActivity/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class Fruits(ref='1');model HeartDiseaseorAttack = Fruits/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class Veggies(ref='1');model HeartDiseaseorAttack = Veggies/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class HvyAlcoholConsump(ref='0');model HeartDiseaseorAttack = HvyAlcoholConsump/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class AnyHealthcare(ref='1');model HeartDiseaseorAttack = AnyHealthcare/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class NoDocbcCost(ref='0');model HeartDiseaseorAttack = NoDocbcCost/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class new\_GenHlth(ref='3');model HeartDiseaseorAttack = new\_GenHlth/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending;class new\_PhysHlth(ref='0'); model HeartDiseaseorAttack = new\_PhysHlth/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending;class new\_Menthlth(ref='0'); model HeartDiseaseorAttack = new\_Menthlth/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class DiffWalk(ref='0');model HeartDiseaseorAttack = DiffWalk/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class Higher\_Education(ref='1');model HeartDiseaseorAttack = Higher\_Education/ RSQUARE clodds=wald orpvalue;run;*  *proc logistic data=project.heart descending; class High\_Earner(ref='1');model HeartDiseaseorAttack = High\_Earner/ RSQUARE clodds=wald orpvalue;run;*  */\*--------------------------------------------------------\*/*  */\*multivariate\*/*  */\*full model\*/*  *proc logistic data=project.heart descending;*  *class*  *Sex(ref='0') AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = Sex AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *Sex(ref='0') AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = Sex AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *Sex(ref='0') AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = Sex AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *run;*  */\*reduced model\*/*  *proc logistic data=project.heart descending;*  *class*  *Sex(ref='0') AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = Sex AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *roc;*  *run;*  */\*--------------------------------------------------------\*/*  */\*Male\*/*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where SEX = 1;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where SEX = 1;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where SEX = 1;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0')Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1')*  *NoDocbcCost(ref='0') new\_GenHlth(ref='3') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck Smoker Stroke new\_Diabetes PhysActivity NoDocbcCost new\_GenHlth*  *DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *Where SEX = 1;*  *roc;*  *run;*  */\*Female\*/*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where SEX = 0;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where SEX = 0;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where SEX = 0;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') new\_PhysHlth(ref='0')*  *NoDocbcCost(ref='0') new\_GenHlth(ref='3') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes NoDocbcCost new\_GenHlth new\_PhysHlth DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *Where SEX = 0;*  *roc;*  *run;*  */\*<30's\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 1;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 1;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 1;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *HighBP(ref='0') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') DiffWalk(ref='0');*  *model*  *HeartDiseaseorAttack = HighBP Smoker Stroke new\_Diabetes new\_GenHlth new\_PhysHlth DiffWalk*  */ RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 1;*  *roc;*  *run;*  */\*30's\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 2;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 2;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 2;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *HighBP(ref='0') HighChol(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = HighBP HighChol BMI\_INDEX Smoker Stroke NoDocbcCost new\_GenHlth new\_PhysHlth DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 2;*  *roc;*  *run;*  */\*40's\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 3;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 3;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 3;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') Fruits(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol Smoker Stroke new\_Diabetes Fruits NoDocbcCost new\_GenHlth new\_PhysHlth DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 3;*  *roc;*  *run;*  */\*50's\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 4;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 4;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 4;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity NoDocbcCost new\_GenHlth new\_PhysHlth DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 4;*  *roc;*  *run;*  */\*60's\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 5;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 5;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 5;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') NoDocbcCost(ref='0') new\_GenHlth(ref='3') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck Smoker Stroke new\_Diabetes NoDocbcCost new\_GenHlth DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 5;*  *roc;*  *run;*  */\*70's\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 6;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 6;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 6;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') DiffWalk(ref='0') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes Veggies NoDocbcCost new\_GenHlth new\_PhysHlth DiffWalk High\_Earner*  */ RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 6;*  *roc;*  *run;*  */\*70's\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = stepwise RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 7;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = forward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 7;*  *run;*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1')*  *Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth*  *new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner*  */ selection = backward RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 7;*  *run;*  */\*reduced\*/*  *proc logistic data=project.heart descending;*  *class*  *SEX(ref='0') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') Fruits(ref='1') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') DiffWalk(ref='0');*  *model*  *HeartDiseaseorAttack = SEX HighBP HighChol CholCheck Smoker Stroke new\_Diabetes Fruits new\_GenHlth new\_PhysHlth DiffWalk*  */ RSQUARE clodds=wald orpvalue;*  *Where AGE\_INDEX = 7;*  *roc;*  *run;*  */\*proc logistic data=project.heart descending; \*/*  */\*class Sex(ref='0') AGE\_INDEX(ref='1') HighBP(ref='0') HighChol(ref='0') CholCheck(ref='0') BMI\_INDEX(ref='2') Smoker(ref='0') Stroke(ref='0') new\_Diabetes(ref='0') PhysActivity(ref='1') Fruits(ref='1') \*/*  */\*Veggies(ref='1') NoDocbcCost(ref='0') new\_GenHlth(ref='3') new\_PhysHlth(ref='0') new\_Menthlth(ref='0') DiffWalk(ref='0') Higher\_Education(ref='1') High\_Earner(ref='1'); \*/*  */\*model HeartDiseaseorAttack(event='1') = Sex AGE\_INDEX HighBP HighChol CholCheck BMI\_INDEX Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_GenHlth \*/*  */\*new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner/ \*/*  */\*selection = s sls = 0.01 ctable; \*/*  */\*roc; \*/*  */\*score data = project.heart out = test; \*/*  */\*run;\*/*  */\*\*/*  */\*\*/*  */\*proc freq data=test; \*/*  */\*tables HeartDiseaseorAttack \* I\_HeartDiseaseorAttack / nocol; \*/*  */\*run; \*/*  */\*Building the model\*/*  *proc hpforest data = project.heart;*  *input Sex / level = nominal;*  *input HighBP HighChol CholCheck Smoker Stroke new\_Diabetes PhysActivity Fruits Veggies NoDocbcCost new\_PhysHlth new\_Menthlth DiffWalk Higher\_Education High\_Earner /level = BINARY;*  *input AGE\_INDEX BMI\_INDEX new\_GenHlth / level = interval;*  *target HeartDiseaseorAttack / level = BINARY;*  *ods output VariableImportance = Variable\_Importance;*  *ods output FitStatistics = fit\_at\_runtime;*  *ods output Baseline = Baseline;*  *save file = "model";*  *run;*  */\*Variable importance\*/*  *title "Feature Importance HeartDisease or Attack";*  *proc sgplot data = Variable\_Importance;*  *vbar Variable /response=Gini groupdisplay = cluster categoryorder=respdesc;*  *run;*  *title "The Average Square Error for final model";*  *proc sgplot data = fit\_at\_runtime;*  *series x=NTrees y=PredAll/legendlabel='Train Error';*  *series x=NTrees y=PredOOB/legendlabel='OOB Error';*  *xaxis values=(0 to 50 by 1);*  *yaxis values=(0 to 0.3 by 0.05) label='Average Square Error';*  *run;*  *title "Missclassfication Error for final model";*  *proc sgplot data=fit\_at\_runtime;*  *series x=NTrees y=MiscAll/legendlabel='Train Error';*  *series x=NTrees y=MiscOob/legendlabel='OOB Error';*  *xaxis values=(0 to 50 by 1);*  *yaxis values=(0 to 0.3 by 0.05) label='Missclassfication Error';*  *run;* |
| --- |