PSTAT 131 Final Project Data Memo

Tao Wang

2022-04-10

Final project data memo

- "According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol." (From https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease) In fact, not only in the United States, a large number of people die of heart disease all over the world every year. Therefore, I choose to study some key indicators of heart disease to help people better understand and prevent heart disease.

An overview of my dataset

- This dataset is from Kaggle (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease). Originally, the dataset come from the 2020 annual CDC survey data of 400k adults related to their health status. However, Kamil Pytlak who cleanded the orginal dataset, and selected the most relevant variables from it in order to help us to do the machine learning projects. Here is the code, and you will see what the first few rows of the dataset look like.

```
# read the the dataset
new_records<- read.csv("heart_2020_cleaned.csv")
head(new_records)
```

```
##   HeartDisease   BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth
## 1           No 16.60     Yes              No     No              3           30
## 2           No 20.34      No              No    Yes              0            0
## 3           No 26.58     Yes              No     No             20           30
## 4           No 24.21      No              No     No              0            0
## 5           No 23.71      No              No     No             28            0
## 6          Yes 28.87     Yes              No     No              6            0
##   DiffWalking    Sex AgeCategory  Race Diabetic PhysicalActivity GenHealth
## 1          No Female       55-59 White      Yes              Yes Very good
## 2          No Female 80 or older White       No              Yes Very good
## 3          No   Male       65-69 White      Yes              Yes      Fair
## 4          No Female       75-79 White       No               No      Good
## 5         Yes Female       40-44 White       No              Yes Very good
## 6         Yes Female       75-79 Black       No               No      Fair
##   SleepTime Asthma KidneyDisease SkinCancer
## 1         5    Yes            No        Yes
## 2         7     No            No         No
## 3         8    Yes            No         No
## 4         6     No            No        Yes
## 5         8     No            No         No
## 6        12     No            No         No
```

```
# check whether there is a missing data
sum(is.na(new_records))
```

```
## [1] 0
```

- This dataset contains 319795 observations with 18 variables (9 booleans, 5 strings and 4 decimals). HeartDisease is the response variable. Other 17 variables are our predictors. I will working with boolean, string, numeric type data. For the result of our code, there is no missing value.

An overview of my research question(s)

- For the purpose of our project, we want to find which variables have a significant effect on the likelihood of heart disease. In order to do it, we should analyze all the predictors (BMI, Smoking, . . .) since all these predictors may have an impact on heart disease. The response variable is HeartDisease, which can tell us whether the respondents had heart disease ("Yes" - respondent had heart disease; "No" - respondent had no heart disease). I think these question will be best answered with regression approach. From what I've studied in medicine, I think the predictors BMI, Smoking, AlcoholDrinking, Stroke, SleepTime will be especially useful. Since we want to find which variables have a significant effect on the likelihood of heart disease, then we can know that the goal of our model is inferential. (It means we want to find the the relationship between the response variable and each predictor, and then we can find some predictors have a significant effect on the likelihood of heart disease.)

My proposed project timeline This is week2, and I have chosen my data set. I will focus on my final project on every Friday (since I can go to the office hour and get help.)

In week 3, I will load my data set (my sure which data set I can used https://www.kaggle.com/datasets/alex teboul/heart-disease-health-indicators-dataset or https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease), and I will convert the booleans and string variables into numeric variable to help us analyze the data.

In week 4, I will complete data cleaning.

In week 5, I will complete data Split.

In week 6 and 7, I will complete exploratory data analysis.

In week 8, I will completed model building.

In week 9, I will complete conclusion (and other explanation part of the project), and I will also check the grammar mistakes. And, get suggestions from professor and TA to improve the project.

In week 10, I will revise my project based on the suggestions from professor and TA, and double check my project with my professor and TA. Do a final check, submit it!

Questions or concerns

- Are there any problems or difficult aspects of the project you anticipate? Yes, I don't know how to do a inference meachine learning project at this time since we haven't learned it. And, when I read the instuction from kaggle (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease), it states, 'note that classes are not balanced, so the classic model application approach is not advisable. Fixing the weights/undersampling should yield significantly betters results.' I don't know what does it means actually.
- Any specific questions you have for me/the instructional team? I wonder whether could you allow us to make a personal appointment with you at the week 9 and week 10 in order to help students to check their project. There are so many students at office hours and some students may not attend the office hours because of courses' schedule, and some students can not get help from you at office hours. So, I wonder if you can provide students with personal meeting in order to help their project. If you could, it may be super helpful.