ISYS621 Project

Chongjia Xu, Huiyi Huang

04/13/2022

Professor Mark Dunn

**Data Brief Background**

The original dataset originated with the Centers for Disease Control and Prevention (CDC) and is a key component of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts yearly telephone surveys to collect data on the health condition of U.S. people. According to the CDC. BRFSS conducts over 400,000 adult interviews each year, making it the world's biggest continually performed health survey system. The survey contains questions like: "Do you have serious difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes in your entire life?". Some of the factors directly or indirectly affect the chance people may have heart disease. The original dataset consists of 401,958 rows and 279 columns. After being cleaned by the Kaggle user, [Kamil Pytlak](https://www.kaggle.com/kamilpytlak), the data has been reduced to 20 variables. This dataset now is more suitable to apply a variety of machine learning approaches, binary classification will be most likely to apply. The file is in csv format and the data is recently updated from 2020. It has 319,794 rows and 18 columns. 9 Boolean, 5 String, and 4 Decimal are included in the dataset.

**Data Name and Definition**

HeartDisease: People who has coronary heart disease (CHD) or myocardial infarction (MI)

BMI: Body Mass Index body, calculated by body mass divided by the square of the body height

Smoking: People who smoked more than 100 cigarettes in whole life answer Yes

AlcoholDring: People who drink more than 14 drinks(alcohol) per week answer Yes

Stroke: Answer Yes if actually had a stroke or told by a doctor

Physical Health: The length of days during past 30 days if had physical illness and injury

MentalHealth: The length of days that feeling negative or bad emotion

DiffWalking: If people feel difficulty during walking or climbing stairs, then answer Yes

Gender: Female or Male

Age Category: Separate by 5 years

Race: The human race of the people

Diabetic: Answer Yes if actually had a diabetic or told by a doctor

PhysicalActivity: Doing physical activity or exercise during previous 30 days, then answer Yes

GenHealth: General Health separates into different category based on the feeling by adults

SleepTime: The length of sleeping that people get in previous 24 hours

Asthma: Answer Yes if actually had a Asthma or told by a doctor

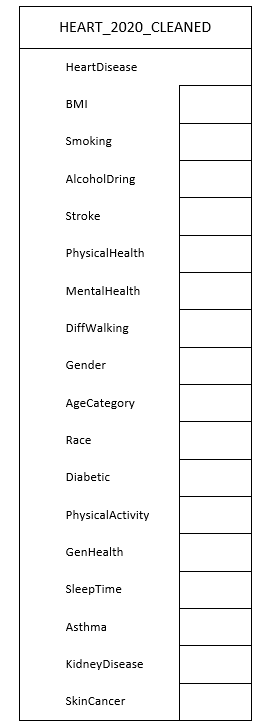
KidneyDisease: Answer Yes if you have kidney stones, bladder infection, incontinence, etc

SkinCancer: Answer Yes if actually had a skin cancer or told by a doctor

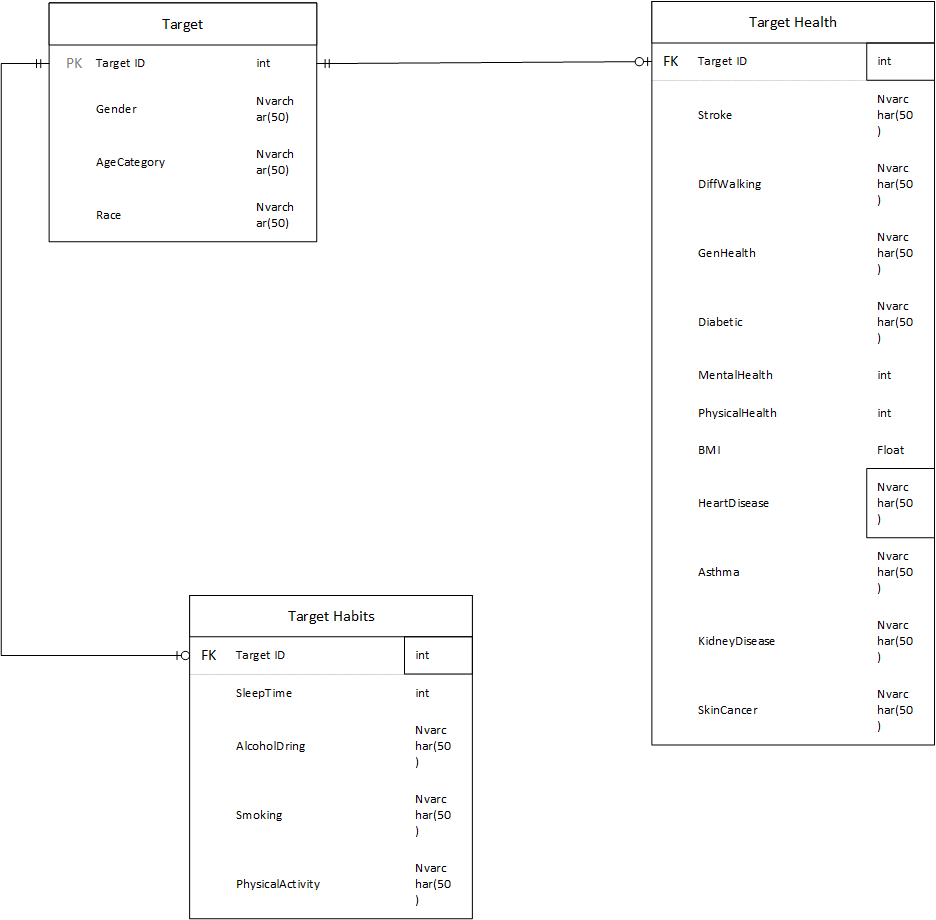
**Metadata-Table**

| **Data Item Metadata** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Min** | **Max** | **Description** | **Source** |
| HeartDisease | Boolean | 2 |  |  | Heart disease status | Heart disease  table |
| BMI | Decimal | 5 | 12.02 | 94.85 | BMI number | Heart disease  table |
| Smoking | Boolean | 2 |  |  | Smoking status | Heart disease  table |
| AlcoholDring | Boolean | 2 |  |  | Drinking status | Heart disease  table |
| Stroke | Boolean | 2 |  |  | Stroke status | Heart disease  table |
| PhysicalHealth | Integer | 2 | 0 | 30 | Physical health grade | Heart disease  table |
| MentalHealth | Integer | 2 | 0 | 30 | Mental health grade | Heart disease  table |
| DiffWalking | Boolean | 2 |  |  | Walking status | Heart disease  table |
| Sex | Alphanumeric | 6 |  |  | Sex | Heart disease  table |
| AgeCategory | Alphanumeric | 30 |  |  | Age | Heart disease  table |
| Race | Alphanumeric | 30 |  |  | Race | Heart disease  table |
| Diabetic | Boolean | 2 |  |  | Diabetic status | Heart disease  table |
| PhysicalActivity | Boolean | 2 |  |  | PhysicalActivity score | Heart disease  table |
| GenHealth | Alphanumeric | 9 |  |  | General Health score | Heart disease  table |
| SleepTime | Integer | 2 | 1 | 24 | Sleep length | Heart disease  table |
| Asthma | Boolean | 2 |  |  | Asthma status | Heart disease  table |
| KidneyDisease | Boolean | 2 |  |  | KidneyDisease status | Heart disease  table |
| SkinCancer | Boolean | 2 |  |  | SkinCancer status | Heart disease  table |

**Current Logic Model**



**Normalized Logical Model (3NF)**

****

**Business Questions**

Which gender/sex has a higher percentage of heart disease rate?

Which race has the highest heart disease?

Does the heart disease rate increase based on the age increase?