# Diabetes Risk Predictions

Data Understanding

Seif Kungulio

07/11/2025

## Evaluating Dataset

## Data Dictionary

## Exploratory Data Analysis

### Data Loading

Load the dataset and name it Diabetes.df

```
Diabetes.df <- read.csv("diabetes_health_indicators_BRFSS2015.csv", header = TRUE)
```

Examine the dimension of the data frame

```
dim(Diabetes.df)
```

```
## [1] 253680     22
```

Display the first six rows of the data frame

```
head(Diabetes.df)
```

```
##   Diabetes_binary HighBP HighChol CholCheck BMI Smoker Stroke
## 1               0      1        1         1  40      1      0
## 2               0      0        0         0  25      1      0
## 3               0      1        1         1  28      0      0
## 4               0      1        0         1  27      0      0
## 5               0      1        1         1  24      0      0
## 6               0      1        1         1  25      1      0
##   HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump
## 1                    0            0      0       1                 0
## 2                    0            1      0       0                 0
## 3                    0            0      1       0                 0
## 4                    0            1      1       1                 0
## 5                    0            1      1       1                 0
## 6                    0            1      1       1                 0
##   AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
## 1             1           0       5       18       15        1   0   9
```

```
## 2             0          1       3        0        0        0   0   7
## 3             1          1       5       30       30        1   0   9
## 4             1          0       2        0        0        0   0  11
## 5             1          0       2        3        0        0   0  11
## 6             1          0       2        0        2        0   1  10
##   Education Income
## 1         4      3
## 2         6      1
## 3         4      8
## 4         3      6
## 5         5      4
## 6         6      8
```

Display the last six rows of the data frame

```
tail(Diabetes.df)
```

```
##        Diabetes_binary HighBP HighChol CholCheck BMI Smoker Stroke
## 253675               0      0        0         1  27      0      0
## 253676               0      1        1         1  45      0      0
## 253677               1      1        1         1  18      0      0
## 253678               0      0        0         1  28      0      0
## 253679               0      1        0         1  23      0      0
## 253680               1      1        1         1  25      0      0
##        HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump
## 253675                    0            0      0       1                 0
## 253676                    0            0      1       1                 0
## 253677                    0            0      0       0                 0
## 253678                    0            1      1       0                 0
## 253679                    0            0      1       1                 0
## 253680                    1            1      1       0                 0
##        AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
## 253675             1           0       1        0        0        0   0   3
## 253676             1           0       3        0        5        0   1   5
## 253677             1           0       4        0        0        1   0  11
## 253678             1           0       1        0        0        0   0   2
## 253679             1           0       3        0        0        0   1   7
## 253680             1           0       2        0        0        0   0   9
##        Education Income
## 253675         6      5
## 253676         6      7
## 253677         2      4
## 253678         5      2
## 253679         5      1
## 253680         6      2
```

Display the structure of the data frame

```
str(Diabetes.df)
```

```
## 'data.frame':    253680 obs. of  22 variables:
##  $ Diabetes_binary     : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ HighBP              : num  1 0 1 1 1 1 1 1 1 0 ...
```

```
##  $ HighChol            : num  1 0 1 0 1 1 0 1 1 0 ...
##  $ CholCheck           : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ BMI                 : num  40 25 28 27 24 25 30 25 30 24 ...
##  $ Smoker              : num  1 1 0 0 0 1 1 1 1 0 ...
##  $ Stroke              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HeartDiseaseorAttack: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ PhysActivity        : num  0 1 0 1 1 1 0 1 0 0 ...
##  $ Fruits              : num  0 0 1 1 1 1 0 0 1 0 ...
##  $ Veggies             : num  1 0 0 1 1 1 0 1 1 1 ...
##  $ HvyAlcoholConsump   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AnyHealthcare       : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ NoDocbcCost         : num  0 1 1 0 0 0 0 0 0 0 ...
##  $ GenHlth             : num  5 3 5 2 2 2 3 3 5 2 ...
##  $ MentHlth            : num  18 0 30 0 3 0 0 0 30 0 ...
##  $ PhysHlth            : num  15 0 30 0 0 2 14 0 30 0 ...
##  $ DiffWalk            : num  1 0 1 0 0 0 0 1 1 0 ...
##  $ Sex                 : num  0 0 0 0 0 1 0 0 0 1 ...
##  $ Age                 : num  9 7 9 11 11 10 9 11 9 8 ...
##  $ Education           : num  4 6 4 3 5 6 6 4 5 4 ...
##  $ Income              : num  3 1 8 6 4 8 7 4 1 3 ...
```

## Data Cleaning

## EDA: Visualizing Data

## EDA: Determining Relationship