

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
```

Preview the data frame

```
glimpse(Diabetes.df)
```

```
## Rows: 253,680
## Columns: 22
## $ Diabetes_binary    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0~
## $ HighBP             <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1~
## $ HighChol           <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1~
## $ CholCheck          <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ BMI                <dbl> 40, 25, 28, 27, 24, 25, 30, 25, 30, 24, 25, 34, 2~
## $ Smoker             <dbl> 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0~
## $ Stroke             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ HeartDiseaseorAttack <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ PhysActivity       <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1~
## $ Fruits             <dbl> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1~
## $ Veggies            <dbl> 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1~
## $ HvyAlcoholConsump  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AnyHealthcare      <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
## $ NoDocbcCost      <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ GenHlth          <dbl> 5, 3, 5, 2, 2, 2, 3, 3, 5, 2, 3, 3, 3, 4, 4, 2, 3~
## $ MentHlth         <dbl> 18, 0, 30, 0, 3, 0, 0, 0, 30, 0, 0, 0, 0, 0, 30, ~
## $ PhysHlth         <dbl> 15, 0, 30, 0, 0, 2, 14, 0, 30, 0, 0, 30, 15, 0, 2~
## $ DiffWalk         <dbl> 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0~
## $ Sex              <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0~
## $ Age              <dbl> 9, 7, 9, 11, 11, 10, 9, 11, 9, 8, 13, 10, 7, 11, ~
## $ Education         <dbl> 4, 6, 4, 3, 5, 6, 6, 4, 5, 4, 6, 5, 5, 4, 6, 6, 4~
## $ Income           <dbl> 3, 1, 8, 6, 4, 8, 7, 4, 1, 3, 8, 1, 7, 6, 2, 8, 3~
```

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 ...
```

Basic Summary and Structure

Summarize the dataset

```
summary(Diabetes.df)
```

```
## Diabetes_binary      HighBP      HighChol      CholCheck
## Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
## 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000
## Median :0.0000   Median :0.000   Median :0.0000   Median :1.0000
## Mean   :0.1393   Mean    :0.429   Mean    :0.4241   Mean    :0.9627
## 3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
## Max.   :1.0000   Max.    :1.000   Max.    :1.0000   Max.    :1.0000
##      BMI      Smoker      Stroke      HeartDiseaseorAttack
## Min.   :12.00   Min.   :0.0000   Min.   :0.000000   Min.   :0.000000
## 1st Qu.:24.00   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.000000
## Median :27.00   Median :0.0000   Median :0.000000   Median :0.000000
## Mean   :28.38   Mean    :0.4432   Mean    :0.04057    Mean    :0.09419
## 3rd Qu.:31.00   3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:0.000000
## Max.   :98.00   Max.    :1.0000   Max.    :1.000000   Max.    :1.000000
## PhysActivity    Fruits      Veggies      HvyAlcoholConsump
## Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
## 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
## Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000
## Mean   :0.7565   Mean    :0.6343   Mean    :0.8114   Mean    :0.0562
## 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
## Max.   :1.0000   Max.    :1.0000   Max.    :1.0000   Max.    :1.0000
## AnyHealthcare    NoDocbcCost      GenHlth      MentHlth
## Min.   :0.0000   Min.   :0.000000   Min.   :1.000    Min.   : 0.000
## 1st Qu.:1.0000   1st Qu.:0.000000   1st Qu.:2.000    1st Qu.: 0.000
## Median :1.0000   Median :0.000000   Median :2.000    Median : 0.000
## Mean   :0.9511   Mean    :0.08418    Mean    :2.511    Mean    : 3.185
## 3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:3.000    3rd Qu.: 2.000
## Max.   :1.0000   Max.    :1.000000   Max.    :5.000    Max.    :30.000
## PhysHlth      DiffWalk      Sex      Age
## Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.000
## 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 6.000
## Median : 0.000   Median :0.0000   Median :0.0000   Median : 8.000
## Mean   : 4.242   Mean    :0.1682   Mean    :0.4403   Mean    : 8.032
## 3rd Qu.: 3.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:10.000
## Max.   :30.000   Max.    :1.0000   Max.    :1.0000   Max.    :13.000
## Education      Income
## Min.   :1.00   Min.   :1.000
## 1st Qu.:4.00   1st Qu.:5.000
## Median :5.00   Median :7.000
## Mean   :5.05   Mean    :6.054
## 3rd Qu.:6.00   3rd Qu.:8.000
## Max.   :6.00   Max.    :8.000
```

Skim for extended summary

```
#skim(Diabetes.df)
```

Data Cleaning

Check for Class Imbalance (target Variable)

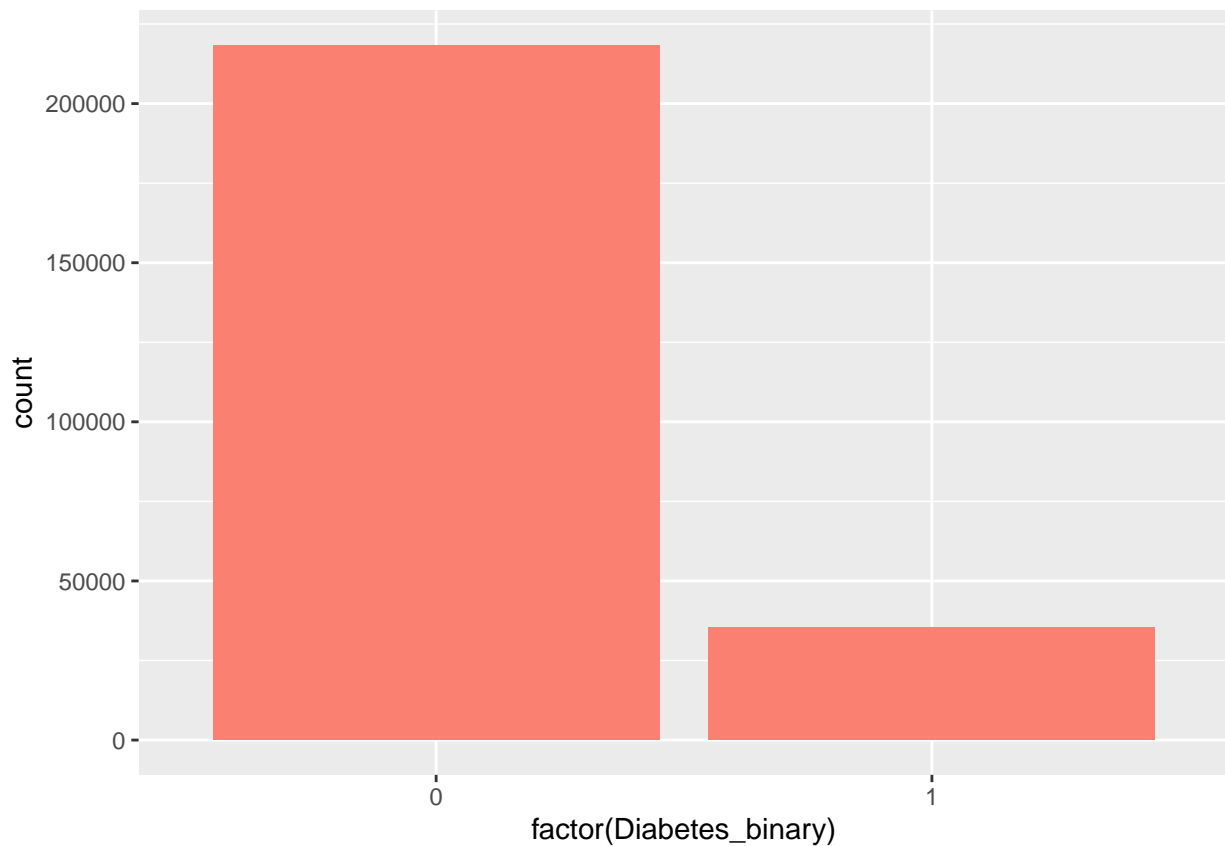
Distribution of diabetes classes

```
Diabetes.df %>%  
  count(Diabetes_binary) %>%  
  mutate(Percent = round(n / sum(n) * 100, 2))
```

```
##  Diabetes_binary      n Percent  
## 1                0 218334   86.07  
## 2                1  35346   13.93
```

Plot the distribution

```
ggplot(Diabetes.df, aes(factor(Diabetes_binary))) +  
  geom_bar(fill = "salmon") +  
  labs()
```



Check for Missing Values and Data Quality

Check for missing values

```
colSums(is.na(Diabetes.df))
```

```
##      Diabetes_binary      HighBP      HighChol
##           0           0           0
##      CholCheck      BMI      Smoker
##           0           0           0
##      Stroke HeartDiseaseorAttack      PhysActivity
##           0           0           0
##      Fruits      Veggies      HvyAlcoholConsump
##           0           0           0
##      AnyHealthcare      NoDocbcCost      GenHlth
##           0           0           0
##      MentHlth      PhysHlth      DiffWalk
##           0           0           0
##      Sex      Age      Education
##           0           0           0
##      Income
##           0
```

EDA: Visualizing Data

Univariate Analysis (Numeric Variables)

EDA: Determining Relationship