Build Prediction Model

Minghe Wang

2025-03-26

Exploratory Data Analysis

```
load("./data/dat1.RData")
load("./data/dat2.RData")

# variable summary
summary(dat1)
```

```
##        id            age            gender          race      smoking
##  Min.   :   1   Min.   :44.00   Min.   :0.0000   1:3221   0:3010
##  1st Qu.:1251   1st Qu.:57.00   1st Qu.:0.0000   2: 278   1:1504
##  Median :2500   Median :60.00   Median :0.0000   3:1036   2: 486
##  Mean   :2500   Mean   :59.97   Mean   :0.4854   4: 465
##  3rd Qu.:3750   3rd Qu.:63.00   3rd Qu.:1.0000
##  Max.   :5000   Max.   :75.00   Max.   :1.0000
##      height          weight           bmi           diabetes
##  Min.   :150.2   Min.   : 56.70   Min.   :18.20   Min.   :0.0000
##  1st Qu.:166.1   1st Qu.: 75.40   1st Qu.:25.80   1st Qu.:0.0000
##  Median :170.1   Median : 80.10   Median :27.60   Median :0.0000
##  Mean   :170.1   Mean   : 80.11   Mean   :27.74   Mean   :0.1544
##  3rd Qu.:174.2   3rd Qu.: 84.90   3rd Qu.:29.50   3rd Qu.:0.0000
##  Max.   :192.9   Max.   :106.00   Max.   :38.80   Max.   :1.0000
##   hypertension        SBP             LDL             time
##  Min.   :0.0000   Min.   :101.0   Min.   : 43.0   Min.   : 30.0
##  1st Qu.:0.0000   1st Qu.:124.0   1st Qu.: 96.0   1st Qu.: 76.0
##  Median :0.0000   Median :130.0   Median :110.0   Median :106.0
##  Mean   :0.4596   Mean   :129.9   Mean   :109.9   Mean   :108.9
##  3rd Qu.:1.0000   3rd Qu.:135.0   3rd Qu.:124.0   3rd Qu.:138.0
##  Max.   :1.0000   Max.   :155.0   Max.   :185.0   Max.   :270.0
##   log_antibody
##  Min.   : 7.765
##  1st Qu.: 9.682
##  Median :10.089
##  Mean   :10.064
##  3rd Qu.:10.478
##  Max.   :11.961
```

```
summary(dat2)
```

```
##        id            age            gender         race     smoking      height
##  Min.   :5001   Min.   :46.00   Min.   :0.000   1:663   0:601   Min.   :149.4
##  1st Qu.:5251   1st Qu.:57.00   1st Qu.:0.000   2: 55   1:296   1st Qu.:166.1
##  Median :5500   Median :60.00   Median :0.000   3:199   2:103   Median :170.2
```

```
##   Mean    :5500   Mean    :60.02   Mean    :0.491   4: 83         Mean    :170.2
##   3rd Qu.:5750   3rd Qu.:63.00   3rd Qu.:1.000                 3rd Qu.:174.2
##   Max.    :6000   Max.    :75.00   Max.    :1.000                 Max.    :190.6
##      weight          bmi           diabetes       hypertension
##   Min.   : 58.80   Min.   :19.80   Min.   :0.000   Min.   :0.000
##   1st Qu.: 75.30   1st Qu.:25.80   1st Qu.:0.000   1st Qu.:0.000
##   Median : 80.20   Median :27.60   Median :0.000   Median :0.000
##   Mean   : 80.13   Mean   :27.72   Mean   :0.157   Mean   :0.456
##   3rd Qu.: 84.40   3rd Qu.:29.60   3rd Qu.:0.000   3rd Qu.:1.000
##   Max.   :101.60   Max.   :35.80   Max.   :1.000   Max.   :1.000
##      SBP             LDL            time          log_antibody
##   Min.   :106.0   Min.   : 46.0   Min.   : 61.0   Min.   : 8.048
##   1st Qu.:124.0   1st Qu.: 96.0   1st Qu.:140.0   1st Qu.: 9.502
##   Median :130.0   Median :112.0   Median :171.0   Median : 9.935
##   Mean   :129.6   Mean   :110.2   Mean   :173.8   Mean   : 9.897
##   3rd Qu.:135.0   3rd Qu.:124.0   3rd Qu.:205.0   3rd Qu.:10.315
##   Max.   :156.0   Max.   :174.0   Max.   :330.0   Max.   :11.852
```

```r
# no missing data
all(is.na(dat1))
```

```
## [1] FALSE
```

```r
all(is.na(dat2))
```

```
## [1] FALSE
```

Model Training