P8106 Final Project: Predicting COVID-19 Recovery Time and Identifying Significant Risk Factors

Runze Cui (rc3521), Yuchen Hua (yh3555), Hongpu Min (hm2946)

2023-05-01

Contents

Background:

Check the report.

Data:

Description check the report.

```
# Dataset Loading
load("data/recovery.Rdata")

set.seed(3521) # Runze Cui's uni(2183):
# Create a first random sample of 2000 participants:
dat1 <- dat[sample(1:10000, 2000),]

set.seed(3555) # Yuchen Hua's uni(3555)
# Create a second random sample of 2000 participants:
dat2 <- dat[sample(1:10000, 2000),]

# Merged the two datasets and remove repeated observations:
dat <- unique(rbind(dat1, dat2))

# Get rid of the id variable from the merged dataset and do the data cleaning:
dat = dat %>%
  select(-id) %>%
  mutate(gender = as.factor(gender)) %>%
  mutate(race = as.factor(race)) %>%
  mutate(smoking = as.factor(smoking)) %>%
  mutate(hypertension = as.factor(hypertension)) %>%
  mutate(diabetes = as.factor(diabetes)) %>%
  mutate(vaccine = as.factor(vaccine)) %>%
  mutate(severity = as.factor(severity)) %>%
  mutate(study = as.factor(study)) %>%
  na.omit() %>%
  relocate(recovery_time)

head(dat)
##      recovery_time age gender race smoking height weight  bmi hypertension
## 8158            52  61      0    1       1  169.9   87.6 30.4            0
## 3387            24  60      1    1       2  173.4   70.6 23.5            0
## 1709            36  60      1    1       1  178.2   79.9 25.1            0
## 4051            23  70      1    4       0  167.4   77.7 27.7            1
## 954             24  63      1    4       0  175.4   88.7 28.8            1
## 531             36  65      0    1       0  160.4   74.4 28.9            1
##      diabetes SBP LDL vaccine severity study
## 8158        0 118 103       0        0     C
## 3387        0 129 101       1        0     B
## 1709        0 130 107       1        0     A
## 4051        0 145 128       1        0     B
## 954         0 131 100       0        0     A
## 531         0 137 153       1        0     A
```

```r
# Separate the data as training and test data:
set.seed(3521)
# Specify rows of training data:
trRows <- createDataPartition(dat$recovery_time, p = 0.7, list = FALSE)

# training data
training <- dat[trRows, ]
## matrix of predictors
x <- model.matrix(recovery_time~.,dat)[trRows,-1]
## vector of response
y <- dat$recovery_time[trRows]

# test data
test <- dat[-trRows, ]
## matrix of predictors
x2 <- model.matrix(recovery_time~.,dat)[-trRows,-1]
## vector of response
y2 <- dat$recovery_time[-trRows]
```