Brain_Stroke

unknownda

2022-10-13

Setting up the data frame

Installing and loading *readr*, *fastDummies*, *caTools*, *dplyr*, *ggplot2*, *gmodels* and *keras* packages. Then import the Brain Stroke data from the *brain_stroke.csv* file. Data Source

The objective is to develop a Neural Network to classify the occcurance of Brain Stroke.

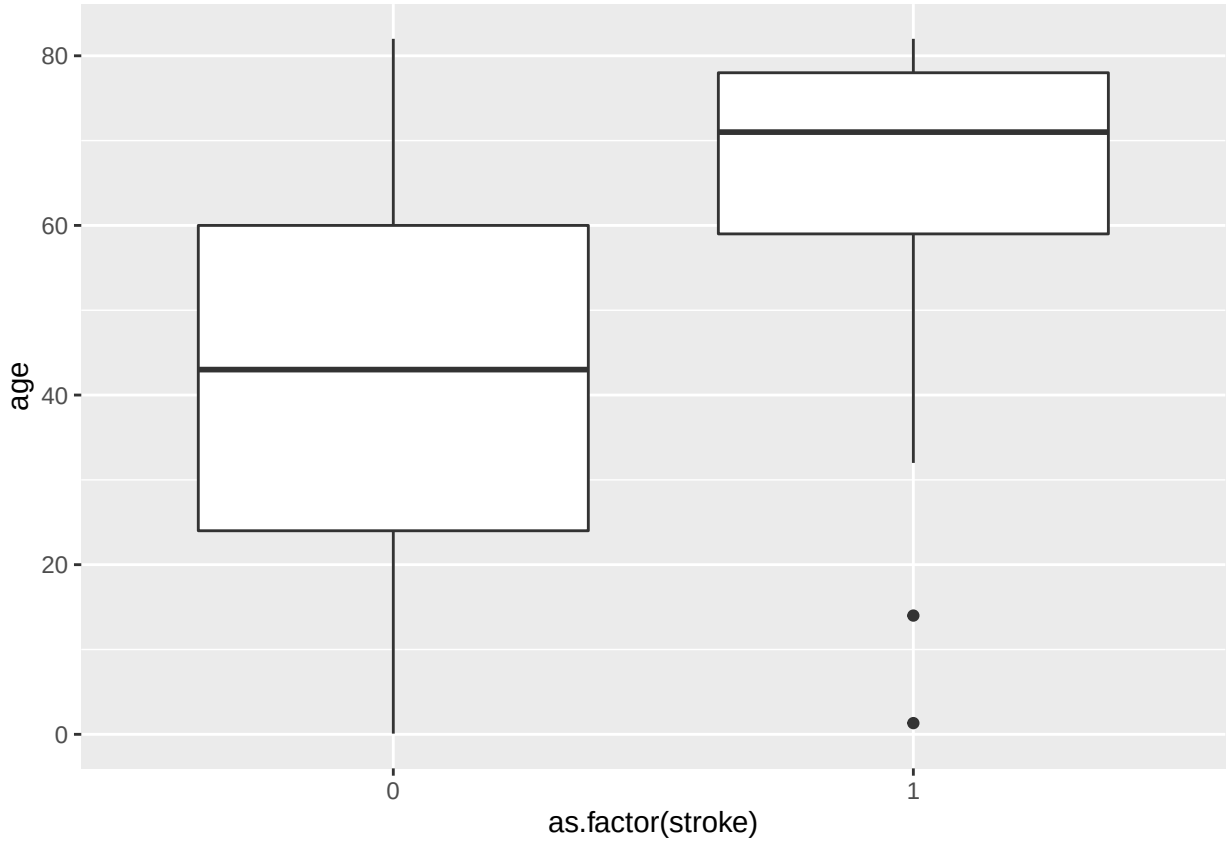
Analyzing Input Data

```
## spec_tbl_df [4,981 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ gender           : chr [1:4981] "Male" "Male" "Female" "Female" ...
##  $ age              : num [1:4981] 67 80 49 79 81 74 69 78 81 61 ...
##  $ hypertension     : num [1:4981] 0 0 0 1 0 1 0 0 1 0 ...
##  $ heart_disease    : num [1:4981] 1 1 0 0 0 1 0 0 0 1 ...
##  $ ever_married     : chr [1:4981] "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr [1:4981] "Private" "Private" "Private" "Self-employed" ...
##  $ Residence_type   : chr [1:4981] "Urban" "Rural" "Urban" "Rural" ...
##  $ avg_glucose_level: num [1:4981] 229 106 171 174 186 ...
##  $ bmi              : num [1:4981] 36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
##  $ smoking_status   : chr [1:4981] "formerly smoked" "never smoked" "smokes" "never smoked" ...
##  $ stroke           : num [1:4981] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   gender = col_character(),
##   ..   age = col_double(),
##   ..   hypertension = col_double(),
##   ..   heart_disease = col_double(),
##   ..   ever_married = col_character(),
##   ..   work_type = col_character(),
##   ..   Residence_type = col_character(),
##   ..   avg_glucose_level = col_double(),
##   ..   bmi = col_double(),
##   ..   smoking_status = col_character(),
##   ..   stroke = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

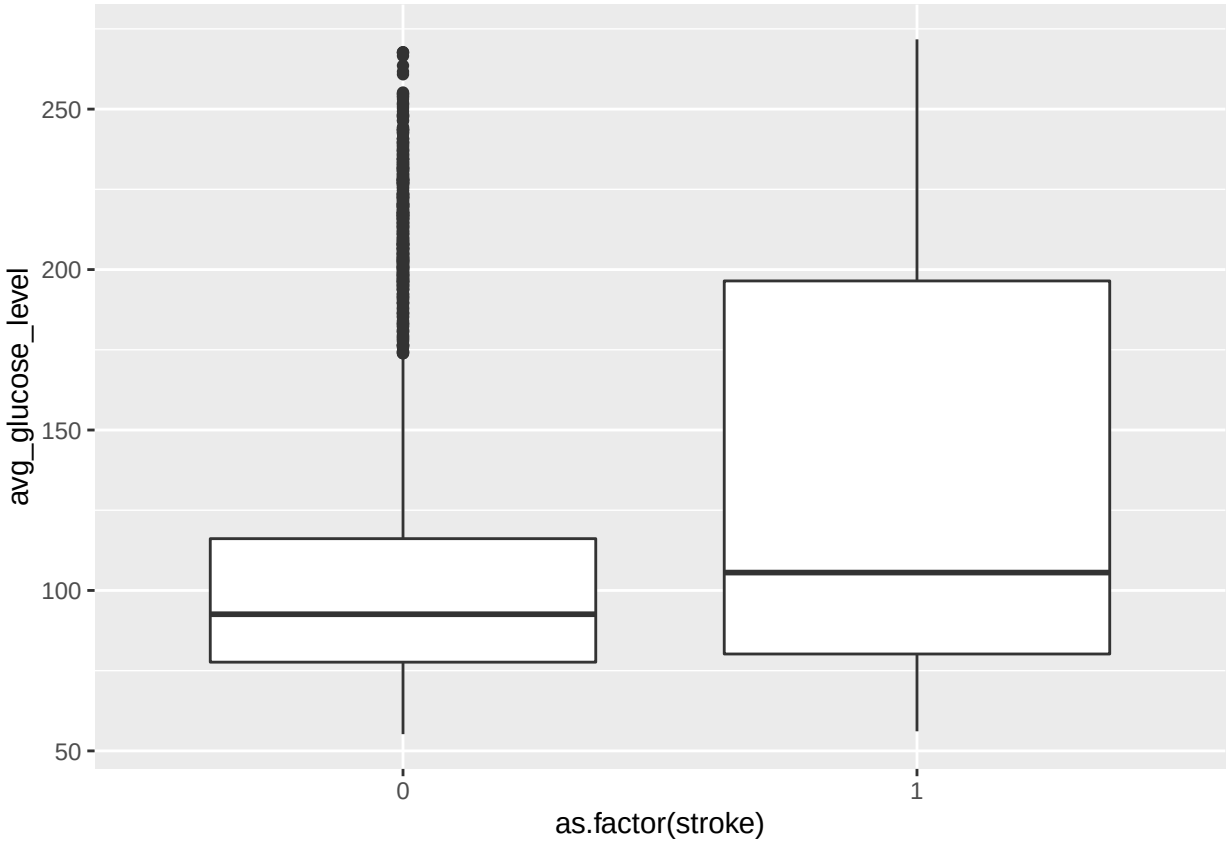
```
##     gender               age         hypertension     heart_disease
##  Length:4981        Min.   : 0.08   Min.   :0.00000   Min.   :0.00000
##  Class :character   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000
##  Mode  :character   Median :45.00   Median :0.00000   Median :0.00000
##                     Mean   :43.42   Mean   :0.09617   Mean   :0.05521
##                     3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000
##                     Max.   :82.00   Max.   :1.00000   Max.   :1.00000
```

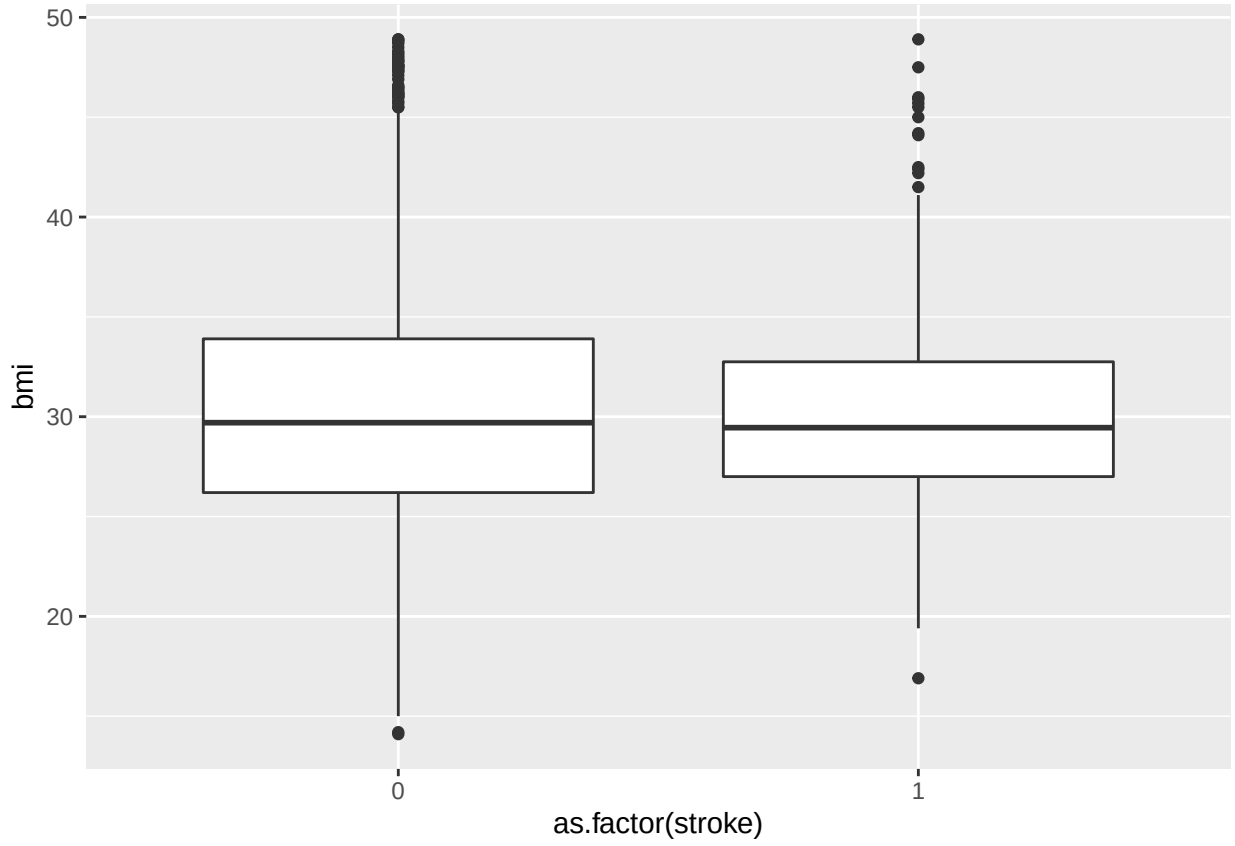
```
##  ever_married        work_type         Residence_type     avg_glucose_level
##  Length:4981        Length:4981        Length:4981        Min.   : 55.12
##  Class :character   Class :character   Class :character   1st Qu.: 77.23
##  Mode  :character   Mode  :character   Mode  :character   Median : 91.85
##                                                           Mean   :105.94
##                                                           3rd Qu.:113.86
##                                                           Max.   :271.74
##       bmi        smoking_status         stroke
##  Min.   :14.0   Length:4981        Min.   :0.00000
##  1st Qu.:23.7   Class :character   1st Qu.:0.00000
##  Median :28.1   Mode  :character   Median :0.00000
##  Mean   :28.5                      Mean   :0.04979
##  3rd Qu.:32.6                      3rd Qu.:0.00000
##  Max.   :48.9                      Max.   :1.00000
```

- The dataset contains no missing values.
- Converting chr columns to factor.



There are no strokes for age < 30 but there are two outliers, so filtering out age < 30.


250
200
150
100
50
avg_glucose_level
0
1
as.factor(stroke)




```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3515
##
##
##                      | df$gender
## as.factor(df$stroke) |    Female |      Male | Row Total |
## ---------------------|-----------|-----------|-----------|
##                    0 |      1944 |      1325 |      3269 |
##                      |     0.595 |     0.405 |     0.930 |
##                      |     0.934 |     0.925 |           |
##                      |     0.553 |     0.377 |           |
## ---------------------|-----------|-----------|-----------|
##                    1 |       138 |       108 |       246 |
##                      |     0.561 |     0.439 |     0.070 |
##                      |     0.066 |     0.075 |           |
##                      |     0.039 |     0.031 |           |
```

```
## ---------------------|-----------|-----------|-----------|
##         Column Total |      2082 |      1433 |      3515 |
##                      |     0.592 |     0.408 |           |
## ---------------------|-----------|-----------|-----------|
##
##
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3515
##
##
##                      | df$hypertension
## as.factor(df$stroke) |         0 |         1 | Row Total |
## ---------------------|-----------|-----------|-----------|
##                    0 |      2864 |       405 |      3269 |
##                      |     0.876 |     0.124 |     0.930 |
##                      |     0.941 |     0.860 |           |
##                      |     0.815 |     0.115 |           |
## ---------------------|-----------|-----------|-----------|
##                    1 |       180 |        66 |       246 |
##                      |     0.732 |     0.268 |     0.070 |
##                      |     0.059 |     0.140 |           |
##                      |     0.051 |     0.019 |           |
## ---------------------|-----------|-----------|-----------|
##         Column Total |      3044 |       471 |      3515 |
##                      |     0.866 |     0.134 |           |
## ---------------------|-----------|-----------|-----------|
##
##
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3515
##
##
##                      | df$heart_disease
```

```
## as.factor(df$stroke) |         0 |         1 | Row Total |
## ---------------------|-----------|-----------|-----------|
##                    0 |      3043 |       226 |      3269 |
##                      |     0.931 |     0.069 |     0.930 |
##                      |     0.939 |     0.828 |           |
##                      |     0.866 |     0.064 |           |
## ---------------------|-----------|-----------|-----------|
##                    1 |       199 |        47 |       246 |
##                      |     0.809 |     0.191 |     0.070 |
##                      |     0.061 |     0.172 |           |
##                      |     0.057 |     0.013 |           |
## ---------------------|-----------|-----------|-----------|
##         Column Total |      3242 |       273 |      3515 |
##                      |     0.922 |     0.078 |           |
## ---------------------|-----------|-----------|-----------|
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3515
##
##
##                      | df$ever_married
## as.factor(df$stroke) |        No |       Yes | Row Total |
## ---------------------|-----------|-----------|-----------|
##                    0 |       373 |      2896 |      3269 |
##                      |     0.114 |     0.886 |     0.930 |
##                      |     0.932 |     0.930 |           |
##                      |     0.106 |     0.824 |           |
## ---------------------|-----------|-----------|-----------|
##                    1 |        27 |       219 |       246 |
##                      |     0.110 |     0.890 |     0.070 |
##                      |     0.068 |     0.070 |           |
##                      |     0.008 |     0.062 |           |
## ---------------------|-----------|-----------|-----------|
##         Column Total |       400 |      3115 |      3515 |
##                      |     0.114 |     0.886 |           |
## ---------------------|-----------|-----------|-----------|
##
##
##
##
##    Cell Contents
## |-------------------------|
```

```
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3515
##
##
##                      | df$work_type
## as.factor(df$stroke) |      Govt_job |       Private | Self-employed |     Row Total |
## ---------------------|---------------|---------------|---------------|---------------|
##                    0 |           552 |          2020 |           697 |          3269 |
##                      |         0.169 |         0.618 |         0.213 |         0.930 |
##                      |         0.944 |         0.932 |         0.915 |               |
##                      |         0.157 |         0.575 |         0.198 |               |
## ---------------------|---------------|---------------|---------------|---------------|
##                    1 |            33 |           148 |            65 |           246 |
##                      |         0.134 |         0.602 |         0.264 |         0.070 |
##                      |         0.056 |         0.068 |         0.085 |               |
##                      |         0.009 |         0.042 |         0.018 |               |
## ---------------------|---------------|---------------|---------------|---------------|
##         Column Total |           585 |          2168 |           762 |          3515 |
##                      |         0.166 |         0.617 |         0.217 |               |
## ---------------------|---------------|---------------|---------------|---------------|
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3515
##
##
##                      | df$Residence_type
## as.factor(df$stroke) |     Rural |     Urban | Row Total |
## ---------------------|-----------|-----------|-----------|
##                    0 |      1603 |      1666 |      3269 |
##                      |     0.490 |     0.510 |     0.930 |
##                      |     0.935 |     0.926 |           |
##                      |     0.456 |     0.474 |           |
## ---------------------|-----------|-----------|-----------|
##                    1 |       112 |       134 |       246 |
##                      |     0.455 |     0.545 |     0.070 |
##                      |     0.065 |     0.074 |           |
```

```
##                       |     0.032 |     0.038 |           |
## ----------------------|-----------|-----------|-----------|
##          Column Total |      1715 |      1800 |      3515 |
##                       |     0.488 |     0.512 |           |
## ----------------------|-----------|-----------|-----------|
##
##
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3515
##
##
##                       | df$smoking_status
## as.factor(df$stroke) | formerly smoked |    never smoked |          smokes |         Unknown |
## ----------------------|-----------------|-----------------|-----------------|-----------------|-------
##                     0 |             706 |            1327 |             604 |             632 |
##                       |           0.216 |           0.406 |           0.185 |           0.193 |
##                       |           0.910 |           0.937 |           0.935 |           0.934 |
##                       |           0.201 |           0.378 |           0.172 |           0.180 |
## ----------------------|-----------------|-----------------|-----------------|-----------------|-------
##                     1 |              70 |              89 |              42 |              45 |
##                       |           0.285 |           0.362 |           0.171 |           0.183 |
##                       |           0.090 |           0.063 |           0.065 |           0.066 |
##                       |           0.020 |           0.025 |           0.012 |           0.013 |
## ----------------------|-----------------|-----------------|-----------------|-----------------|-------
##          Column Total |             776 |            1416 |             646 |             677 |
##                       |           0.221 |           0.403 |           0.184 |           0.193 |
## ----------------------|-----------------|-----------------|-----------------|-----------------|-------
##
##
```

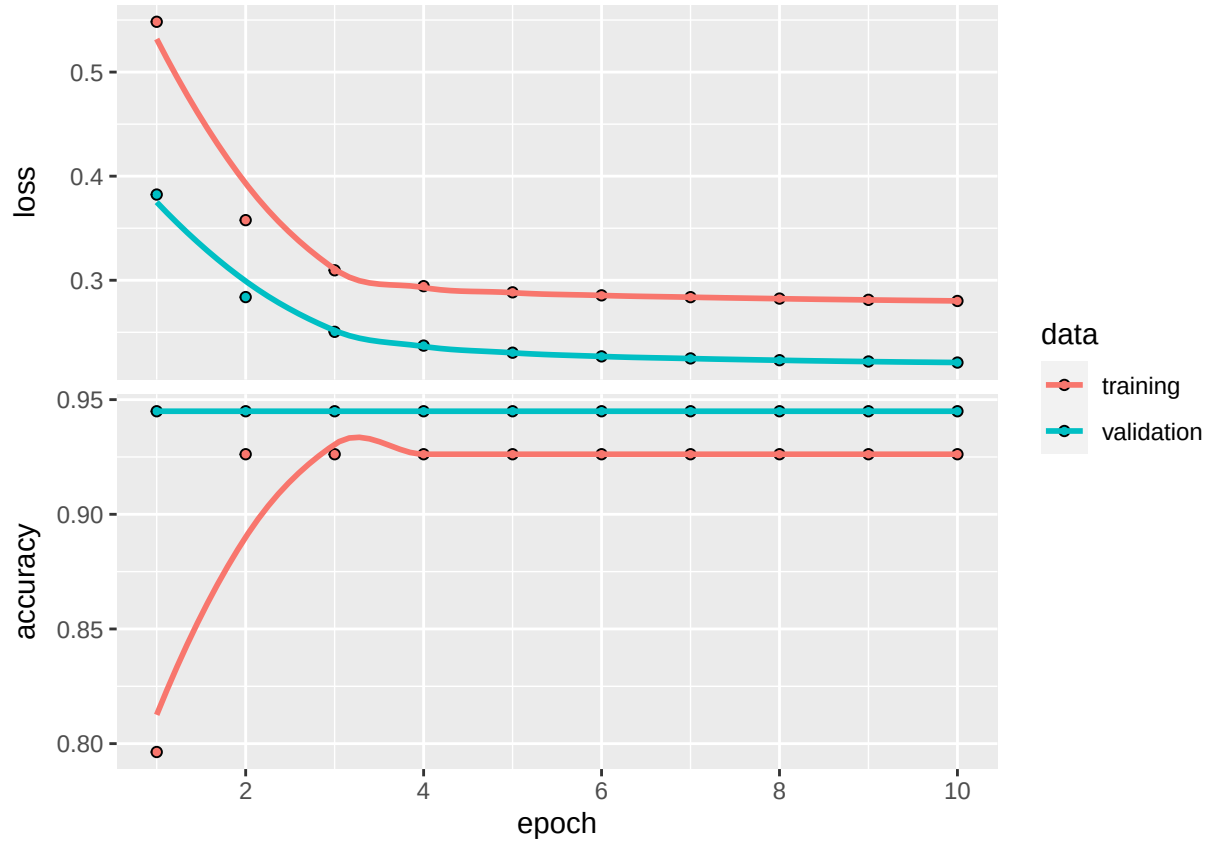
Training & Test Set

```
set.seed(0)
split <- sample.split(df_final$stroke, SplitRatio = 0.8)
training_set <- subset(df_final, split==TRUE)
test_set <- subset(df_final, split==FALSE)
```

Neural Network

```
## Loaded Tensorflow version 2.9.2
```

```
## Model: "sequential"
## _________________________________________________________________________
##  Layer (type)                       Output Shape                    Param #
## =========================================================================
##  flatten (Flatten)                  (None, 14)                      0
##  dense_1 (Dense)                    (None, 5)                       75
##  dense (Dense)                      (None, 1)                       6
## =========================================================================
## Total params: 81
## Trainable params: 81
## Non-trainable params: 0
## _________________________________________________________________________
```




```
##
##  Test Set Performance
```

```
## Test loss: 0.2681147
```

```
## Test Accuracy: 0.9302987
```