Build and deploy a stroke prediction model using R

Aman Singh

2025-06-09

About Data Analysis Report

This RMarkdown file contains the report of the data analysis done for the project on building and deploying a stroke prediction model in R. It contains analysis such as data exploration, summary statistics and building the prediction models. The final report was completed on Mon Jun 9 20:17:12 2025.

Data Description:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This data set is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Task One: Import data and data preprocessing

Load data and install packages

```
options(repos = c(CRAN="https://cran.r-project.org"))
install.packages("tidyverse")
```

```
## Installing package into 'C:/Users/meghn/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\meghn\AppData\Local\Temp\RtmpSqQgkX\downloaded_packages
```

```
install.packages("tidymodels")
```

```
## Installing package into 'C:/Users/meghn/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)

## package 'tidymodels' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\meghn\AppData\Local\Temp\RtmpSqQgkX\downloaded_packages
```

```r
install.packages("dplyr")
```

```
## Installing package into 'C:/Users/meghn/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)

## package 'dplyr' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'dplyr'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\meghn\AppData\Local\R\win-library\4.5\00LOCK\dplyr\libs\x64\dplyr.dll
## to C:\Users\meghn\AppData\Local\R\win-library\4.5\dplyr\libs\x64\dplyr.dll:
## Permission denied

## Warning: restored 'dplyr'

##
## The downloaded binary packages are in
##  C:\Users\meghn\AppData\Local\Temp\RtmpSqQgkX\downloaded_packages
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.3.0 --
## v broom        1.0.8     v rsample      1.3.0
## v dials        1.4.0     v tune         1.3.0
## v infer        1.0.8     v workflows    1.2.0
## v modeldata    1.4.0     v workflowsets 1.1.1
## v parsnip      1.3.2     v yardstick    1.3.2
## v recipes      1.3.1
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
```

```
library(workflows)
library(tune)
library(readr)
install.packages("caret")
```

```
## Installing package into 'C:/Users/meghn/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)
```

```
## package 'caret' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'caret'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\meghn\AppData\Local\R\win-library\4.5\00LOCK\caret\libs\x64\caret.dll
## to C:\Users\meghn\AppData\Local\R\win-library\4.5\caret\libs\x64\caret.dll:
## Permission denied
```

```
## Warning: restored 'caret'
```

```
##
## The downloaded binary packages are in
##  C:\Users\meghn\AppData\Local\Temp\RtmpSqQgkX\downloaded_packages
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
stroke <- read_csv("healthcare-dataset-stroke-data.csv")
```

```
## Rows: 5110 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): gender, ever_married, work_type, Residence_type, bmi, smoking_status
## dbl (6): id, age, hypertension, heart_disease, avg_glucose_level, stroke
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(stroke)
```

Describe and explore the data

```
head(stroke)
```

```
## # A tibble: 6 x 12
##      id gender   age hypertension heart_disease ever_married work_type
##   <dbl> <chr>  <dbl>        <dbl>         <dbl> <chr>        <chr>
## 1  9046 Male      67            0             1 Yes          Private
## 2 51676 Female    61            0             0 Yes          Self-employed
## 3 31112 Male      80            0             1 Yes          Private
## 4 60182 Female    49            0             0 Yes          Private
## 5  1665 Female    79            1             0 Yes          Self-employed
## 6 56669 Male      81            0             0 Yes          Private
## # i 5 more variables: Residence_type <chr>, avg_glucose_level <dbl>, bmi <chr>,
## #   smoking_status <chr>, stroke <dbl>
```

```
glimpse(stroke)
```

```
## Rows: 5,110
## Columns: 12
## $ id                <dbl> 9046, 51676, 31112, 60182, 1665, 56669, 53882, 10434~
## $ gender            <chr> "Male", "Female", "Male", "Female", "Female", "Male"~
## $ age               <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, ~
## $ hypertension      <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1~
## $ heart_disease     <dbl> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0~
## $ ever_married      <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No~
## $ work_type         <chr> "Private", "Self-employed", "Private", "Private", "S~
## $ Residence_type    <chr> "Urban", "Rural", "Rural", "Urban", "Rural", "Urban"~
## $ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.0~
## $ bmi               <chr> "36.6", "N/A", "32.5", "34.4", "24", "29", "27.4", "~
## $ smoking_status    <chr> "formerly smoked", "never smoked", "never smoked", "~
## $ stroke            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
sapply(stroke,class)
```

```
##                id            gender               age      hypertension
##         "numeric"       "character"         "numeric"         "numeric"
##     heart_disease      ever_married         work_type    Residence_type
##         "numeric"       "character"       "character"       "character"
## avg_glucose_level               bmi    smoking_status            stroke
##         "numeric"       "character"       "character"         "numeric"
```

```
clean_stroke <- drop_na(stroke)
sum(is.na(clean_stroke))
```

```
## [1] 0
```

```r
summary(clean_stroke)
```

```
##        id           gender               age         hypertension
##  Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Mode  :character   Median :45.00   Median :0.00000
##  Mean   :36518                      Mean   :43.23   Mean   :0.09746
##  3rd Qu.:54682                      3rd Qu.:61.00   3rd Qu.:0.00000
##  Max.   :72940                      Max.   :82.00   Max.   :1.00000
##  heart_disease     ever_married        work_type         Residence_type
##  Min.   :0.00000   Length:5110        Length:5110        Length:5110
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.05401
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  avg_glucose_level     bmi            smoking_status         stroke
##  Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000
##  1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000
##  Median : 91.89    Mode  :character   Mode  :character   Median :0.00000
##  Mean   :106.15                                          Mean   :0.04873
##  3rd Qu.:114.09                                          3rd Qu.:0.00000
##  Max.   :271.74                                          Max.   :1.00000
```

```r
stroke$gender <- as.factor(stroke$gender)
stroke$ever_married <- as.factor(stroke$ever_married)
stroke$work_type <- as.factor(stroke$work_type)
stroke$Residence_type <- as.factor(stroke$Residence_type)
stroke$smoking_status <- as.factor(stroke$smoking_status)
stroke$hypertension <- as.factor(stroke$hypertension)
stroke$stroke <- as.factor(stroke$stroke)
stroke$heart_disease <- as.factor(stroke$heart_disease)
stroke$bmi <- as.numeric(stroke$bmi)
```

```
## Warning: NAs introduced by coercion
```

```r
stroke[stroke == "Unknown"] <- NA
summary(stroke)
```

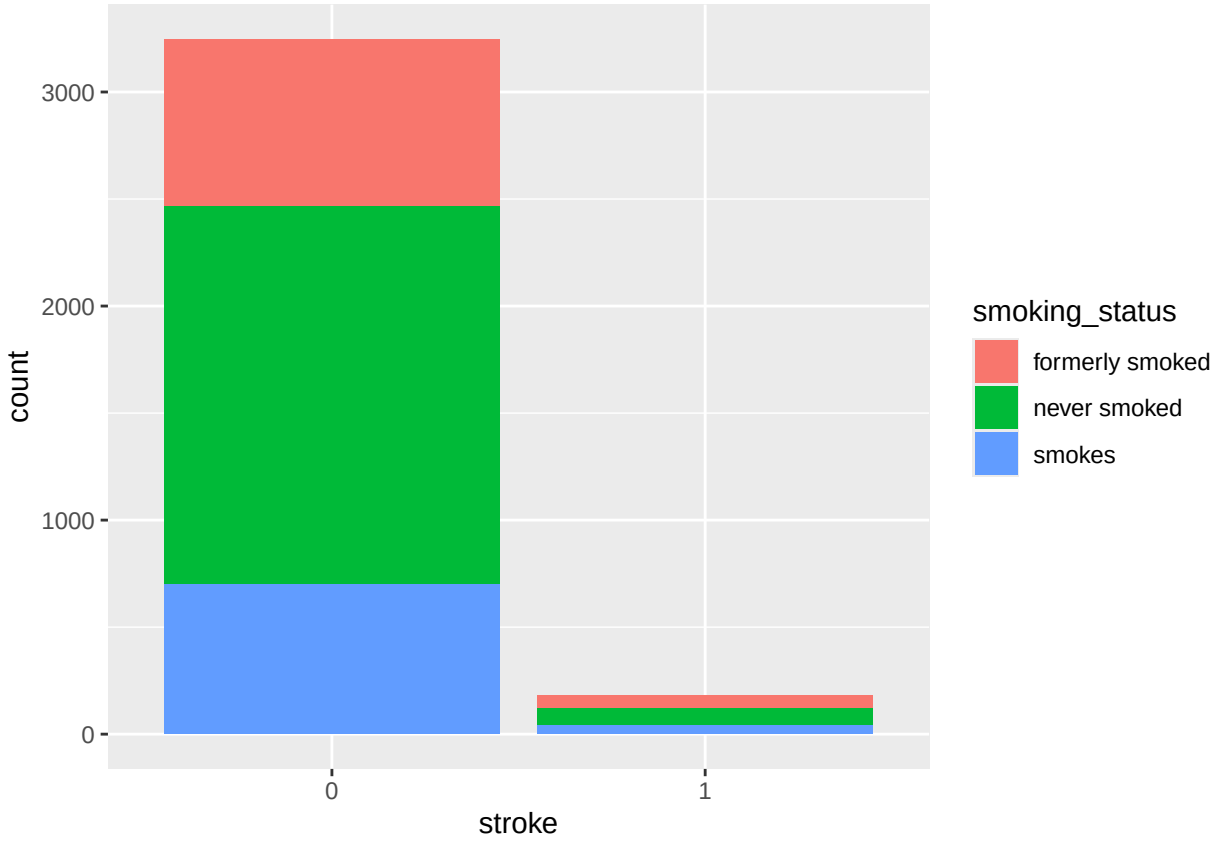
```
##        id           gender          age        hypertension heart_disease
##  Min.   :   67   Female:2994   Min.   : 0.08   0:4612       0:4834
##  1st Qu.:17741   Male  :2115   1st Qu.:25.00   1: 498       1: 276
##  Median :36932   Other :   1   Median :45.00
##  Mean   :36518                 Mean   :43.23
##  3rd Qu.:54682                 3rd Qu.:61.00
##  Max.   :72940                 Max.   :82.00
##
##  ever_married         work_type    Residence_type avg_glucose_level
##  No :1757     children     : 687   Rural:2514     Min.   : 55.12
##  Yes:3353     Govt_job     : 657   Urban:2596     1st Qu.: 77.25
##               Never_worked :  22                  Median : 91.89
##               Private      :2925                  Mean   :106.15
```

```
##                Self-employed: 819                  3rd Qu.:114.09
##                                                    Max.   :271.74
##
##       bmi                smoking_status stroke
##  Min.   :10.30   formerly smoked: 885   0:4861
##  1st Qu.:23.50   never smoked   :1892   1: 249
##  Median :28.10   smokes         : 789
##  Mean   :28.89   Unknown        :   0
##  3rd Qu.:33.10   NA's           :1544
##  Max.   :97.60
##  NA's   :201
```

```
clean_stroke <- drop_na(stroke)
summary(clean_stroke)
```

```
##        id           gender          age        hypertension heart_disease
##  Min.   :   84   Female:2086   Min.   :10.00   0:3018       0:3220
##  1st Qu.:18998   Male  :1339   1st Qu.:34.00   1: 408       1: 206
##  Median :38069   Other :   1   Median :50.00
##  Mean   :37339                 Mean   :48.65
##  3rd Qu.:55464                 3rd Qu.:63.00
##  Max.   :72915                 Max.   :82.00
##  ever_married         work_type    Residence_type avg_glucose_level
##  No : 827     children     :  68   Rural:1681     Min.   : 55.12
##  Yes:2599     Govt_job     : 514   Urban:1745     1st Qu.: 77.24
##               Never_worked :  14                  Median : 92.36
##               Private      :2201                  Mean   :108.32
##               Self-employed: 629                  3rd Qu.:116.21
##                                                   Max.   :271.74
##       bmi                smoking_status stroke
##  Min.   :11.50   formerly smoked: 837   0:3246
##  1st Qu.:25.30   never smoked   :1852   1: 180
##  Median :29.10   smokes         : 737
##  Mean   :30.29   Unknown        :   0
##  3rd Qu.:34.10
##  Max.   :92.00
```

```
ggplot(clean_stroke, aes(x = stroke , fill = smoking_status)) +
  geom_bar() +
  facet_grid()
```



Task Two: Build prediction models

```
stroke_split <- createDataPartition(clean_stroke$stroke, p=0.80, list=FALSE)

stroke_cv <- clean_stroke[-stroke_split,]
stroke_train <- clean_stroke[stroke_split,]
sum(is.na(stroke_train))
```

```
## [1] 0
```

```
sum(is.na(stroke_cv))
```

```
## [1] 0
```

Task Three: Evaluate and select prediction models

```
control <- trainControl(method = "cv", number = 10)
metric <- "Accuracy"
```

```r
m_cart <- train(stroke ~ gender + age + hypertension + heart_disease + ever_married + work_type + avg_gl
m_knn <- train(stroke ~ gender + age + hypertension + heart_disease + ever_married + work_type + avg_glu
m_svm <- train(stroke ~ gender + age + hypertension + heart_disease + ever_married + work_type + avg_glu
```

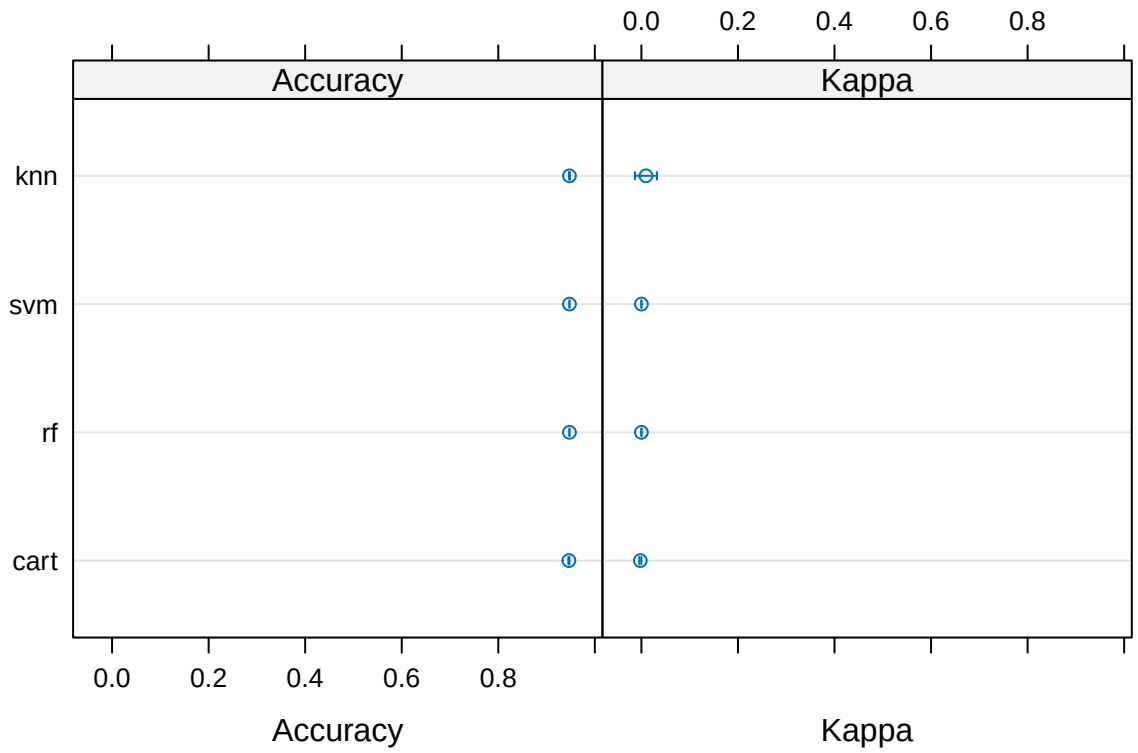
```
## Warning in .local(x, ...): Variable(s) '' constant. Cannot scale data.
```

```r
m_rf <- train(stroke ~ gender + age + hypertension + heart_disease + ever_married + work_type + avg_gluc

results <- resamples(list(cart = m_cart, knn = m_knn, svm = m_svm, rf = m_rf))
summary(results)
```

```
##
## Call:
## summary.resamples(object = results)
##
## Models: cart, knn, svm, rf
## Number of resamples: 10
##
## Accuracy
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## cart 0.9446064 0.9446064 0.9473684 0.9462943 0.9473684 0.9475219    0
## knn  0.9446064 0.9473684 0.9475219 0.9474613 0.9475219 0.9502924    0
## svm  0.9473684 0.9473684 0.9475219 0.9474605 0.9475219 0.9475219    0
## rf   0.9473684 0.9473684 0.9475219 0.9474605 0.9475219 0.9475219    0
##
## Kappa
##              Min.      1st Qu. Median         Mean 3rd Qu.      Max. NA's
## cart -0.005554698 -0.005554698      0 -0.002221879       0 0.0000000    0
## knn  -0.005554698  0.000000000      0  0.009472385       0 0.1002786    0
## svm   0.000000000  0.000000000      0  0.000000000       0 0.0000000    0
## rf    0.000000000  0.000000000      0  0.000000000       0 0.0000000    0
```

```r
dotplot(results)
```



Confidence Level: 0.95

Task Four: Deploy the prediction model

```
print(m_knn)
```

```
## k-Nearest Neighbors
##
## 3426 samples
##    9 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3084, 3083, 3083, 3083, 3084, 3084, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.9430822  0.034434339
##   7  0.9445416  0.012724104
##   9  0.9474613  0.009472385
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

```r
predictions <- predict(m_knn, stroke_cv)
confusionMatrix(predictions, stroke_cv$stroke)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 649  36
##          1   0   0
##
##                Accuracy : 0.9474
##                  95% CI : (0.928, 0.9629)
##     No Information Rate : 0.9474
##     P-Value [Acc > NIR] : 0.5442
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 5.433e-09
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9474
##          Neg Pred Value :    NaN
##              Prevalence : 0.9474
##          Detection Rate : 0.9474
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

Task Five: Findings and Conclusions

The data model has been able to predict the stroke based on nine different predictors. According to my analysis the type of residence does not have significant impact on the stroke prediction. After analysis of all models the SVM model gives the best output. The accuracy is around 94.74% which is under the confidence interval of 95%. The validation data confirms to the probability and thus model predict the stroke correctly.