Exercicio_2

Alex

2025-04-10

```
avc_pgr <- 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.
```

```
avc_pgr <- avc_pgr %>% filter(age >= 18)

avc_pgr <- avc_pgr %>% mutate(bmi=na_if(bmi,"N/A"))%>% mutate(bmi=as.numeric(bmi))
avc_pgr <- avc_pgr %>% select(-id)
avc_pgr <-avc_pgr  %>% filter(gender != "Other")

avc_pgr$gender<-factor(avc_pgr$gender)
avc_pgr$ever_married <- factor(avc_pgr$ever_married)
avc_pgr$work_type <- factor(avc_pgr$work_type)
avc_pgr$Residence_type <- factor(avc_pgr$Residence_type)
avc_pgr$smoking_status <- factor(avc_pgr$smoking_status)

avc_pgr$hypertension <- factor(avc_pgr$hypertension, levels = c(0,1), labels = c("No","Yes"))
avc_pgr$heart_disease <- factor(avc_pgr$heart_disease, levels = c(0,1), labels = c("No","Ye
s"))

avc_pgr$stroke <-factor(avc_pgr$stroke, levels = c(0,1), labels = c("SemAVC","AVC"))
avc_pgr$stroke <- relevel(avc_pgr$stroke, ref = "AVC")
```

```
summary(avc_pgr)
```

```
##     gender          age        hypertension heart_disease ever_married
##  Female:2576   Min.   :18.00   No :3756     No :3978      No : 900
##  Male  :1677   1st Qu.:36.00   Yes: 497     Yes: 275      Yes:3353
##                Median :51.00
##                Mean   :50.21
##                3rd Qu.:64.00
##                Max.   :82.00
##
##          work_type    Residence_type avg_glucose_level      bmi
##  Govt_job     : 651   Rural:2084     Min.   : 55.12    Min.   :11.30
##  Never_worked :   5   Urban:2169     1st Qu.: 77.48    1st Qu.:25.40
##  Private      :2790                  Median : 92.44    Median :29.20
##  Self-employed: 807                  Mean   :108.51    Mean   :30.43
##                                      3rd Qu.:116.12    3rd Qu.:34.20
##                                      Max.   :271.74    Max.   :92.00
##                                                        NA's   :181
##          smoking_status    stroke
##  formerly smoked: 859   AVC   : 247
##  never smoked   :1752   SemAVC:4006
##  smokes         : 780
##  Unknown        : 862
##
##
##
```

```
train_idx <- createDataPartition(avc_pgr$stroke, p=0.7, list=FALSE)
train_data <- avc_pgr[train_idx,]
test_data <- avc_pgr[-train_idx,]
```

```
prop.table(table(test_data$stroke))
```

```
##
##        AVC     SemAVC
## 0.05803922 0.94196078
```

```
prop.table(table(train_data$stroke))
```

```
##
##        AVC     SemAVC
## 0.05809268 0.94190732
```

```
preproc <- preProcess(train_data, method = c("medianImpute","center", "scale"))
```

```
train_processed <- predict(preproc, train_data)
test_processed <- predict(preproc, test_data)
```

```
nzv_indices <- nearZeroVar(train_processed, saveMetrics = TRUE)
nzv_indices
```

```
##                   freqRatio percentUnique zeroVar   nzv
## gender             1.565030    0.06715917   FALSE FALSE
## age                1.106061    2.18267293   FALSE FALSE
## hypertension       7.436261    0.06715917   FALSE FALSE
## heart_disease     13.964824    0.06715917   FALSE FALSE
## ever_married       3.818770    0.06715917   FALSE FALSE
## work_type          3.423818    0.13431833   FALSE FALSE
## Residence_type     1.048143    0.06715917   FALSE FALSE
## avg_glucose_level  1.000000   87.50839490   FALSE FALSE
## bmi                5.692308   12.12222968   FALSE FALSE
## smoking_status     2.069421    0.13431833   FALSE FALSE
## stroke            16.213873    0.06715917   FALSE FALSE
```

```
ctrl <- trainControl(
  method = 'cv',
  number = 5,
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  sampling = 'smote'
)

model_glm <- train(stroke ~ .,
                   data= train_processed,
                   method = 'glm',
                   family = binomial,
                   trControl = ctrl,
                   metric = "ROC"
                   )
```

```
## Loading required package: recipes
```

```
##
## Attaching package: 'recipes'
```

```
## The following object is masked from 'package:stringr':
##
##     fixed
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
print(model_glm)
```

```
## Generalized Linear Model
##
## 2978 samples
##   10 predictor
##    2 classes: 'AVC', 'SemAVC'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 2382, 2382, 2383, 2383, 2382
## Addtional sampling using SMOTE
##
## Resampling results:
##
##   ROC        Sens       Spec
##   0.8016525  0.7452101  0.715508
```

```
glmpedclass <- predict(model_glm, newdata = test_processed)
glmprob <- predict(model_glm, newdata = test_processed, type = 'prob')

preds <- ifelse(glmprob[,"AVC"]> 0.30,"AVC","SemAVC")
preds = factor(preds)

confusionMatrix(preds, test_processed$stroke, positive = 'AVC')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction AVC SemAVC
##     AVC     68    548
##     SemAVC   6    653
##
##                Accuracy : 0.5655
##                  95% CI : (0.5378, 0.5929)
##     No Information Rate : 0.942
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1043
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.91892
##             Specificity : 0.54371
##          Pos Pred Value : 0.11039
##          Neg Pred Value : 0.99090
##              Prevalence : 0.05804
##          Detection Rate : 0.05333
##    Detection Prevalence : 0.48314
##       Balanced Accuracy : 0.73132
##
##        'Positive' Class : AVC
##
```

```
roc_obj <- roc(test_data$stroke, glmprob$AVC)
```

```
## Setting levels: control = AVC, case = SemAVC
```

```
## Setting direction: controls > cases
```

```
plot(roc_obj, col = "blue",main ="ROC curve for stroke prediction")
abline(a=0, b=1, lty =2, col="grey")
```