test

Jenny

10/26/2020

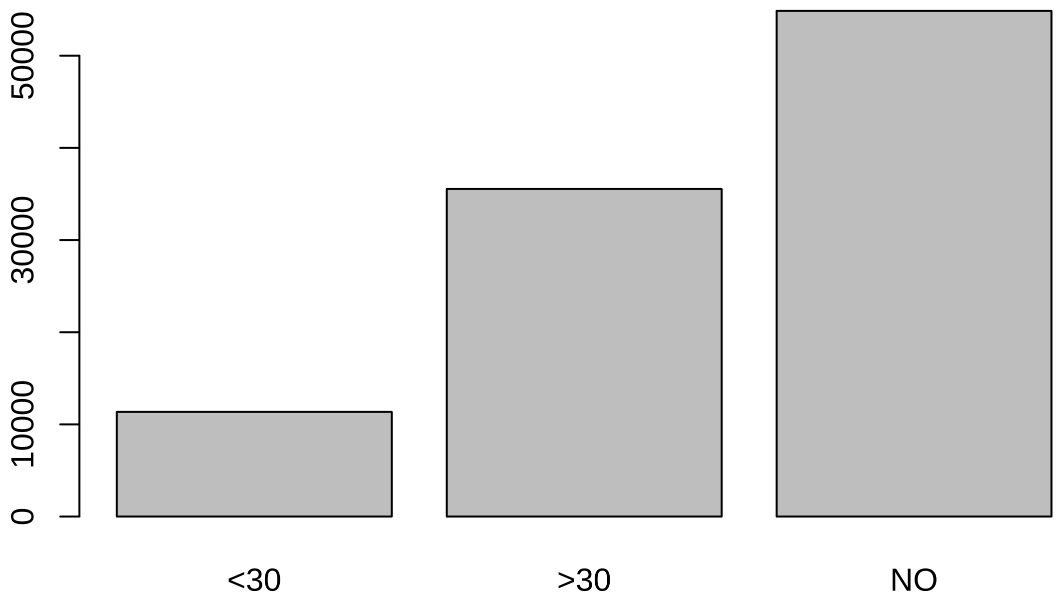
Primero, para la importación de los datos, se eliminan las variables cuya varianza es 0 o casi 0.

```
diabetic_data <- subset(diabetic_data, select=-c(encounter_id,patient_nbr,examide,citoglipton,acetohexam
```

Variable a estimar

La variable a estimar es "readmitted", la cual indica si un paciente ingresado por urgencias con diabetes deberá volver nuevamente antes de 30 días, después de 30 días o ya no debe volver.

```
barplot(table(factor(diabetic_data$readmitted)))
```



Preparación de datos

A continuación se alistarán los datos para el tratamiento de datos nulos.

```
diabetic_data <- replace_with_na(data=diabetic_data,replace=list(admission_type_id = c(5,6,8)))
diabetic_data <- replace_with_na(data=diabetic_data,replace=list(c('?')))
diabetic_data <- replace_with_na(data=diabetic_data,replace=list(gender = c('Unknown/Invalid')))
diabetic_data <- replace_with_na(data=diabetic_data,replace=list(discharge_disposition_id = c(18,25,26)
diabetic_data <- replace_with_na(data=diabetic_data,replace=list(admission_source_id = c(9,15,17,20,21)

factor_to_number = function(x){
```

```r
  if (class(x)=='factor')
    x%>%as.numeric(factor(x))
  else
    x
}
diabetic_data <-as.data.frame(sapply(diabetic_data,factor_to_number))

diabetic_data$admission_type_id <- factor(diabetic_data$admission_type_id)
diabetic_data$race <- factor(diabetic_data$race)
diabetic_data$gender <- factor(diabetic_data$gender)
diabetic_data$discharge_disposition_id <- factor(diabetic_data$discharge_disposition_id)
diabetic_data$admission_source_id <- factor(diabetic_data$admission_source_id)
```

Ahora, realizaremos la imputación de los valores faltantes por medio de la librería **MICE**, utilizando el modelo **predictive mean matching** para imputación de valores de cualquier tipo, en este caso variables categóricas.

```r
mice_diabetic_data <- mice(diabetic_data,m=2,maxit=3,meth='pmm',seed=500)
```

```
##
##  iter imp variable
##   1   1  gender  admission_type_id  discharge_disposition_id  admission_source_id
##   1   2  gender  admission_type_id  discharge_disposition_id  admission_source_id
##   2   1  gender  admission_type_id  discharge_disposition_id  admission_source_id
##   2   2  gender  admission_type_id  discharge_disposition_id  admission_source_id
##   3   1  gender  admission_type_id  discharge_disposition_id  admission_source_id
##   3   2  gender  admission_type_id  discharge_disposition_id  admission_source_id
```

```
## Warning: Number of logged events: 24
```

```r
summary(mice_diabetic_data)
```

```
## Class: mids
## Number of multiple imputations:  2
## Imputation methods:
##                     race                   gender                      age
##                       ""                    "pmm"                       ""
##                   weight        admission_type_id discharge_disposition_id
##                       ""                    "pmm"                    "pmm"
##      admission_source_id         time_in_hospital               payer_code
##                    "pmm"                       ""                       ""
##        medical_specialty       num_lab_procedures           num_procedures
##                       ""                       ""                       ""
##          num_medications        number_outpatient         number_emergency
##                       ""                       ""                       ""
##         number_inpatient                   diag_1                   diag_2
##                       ""                       ""                       ""
##                   diag_3         number_diagnoses            max_glu_serum
##                       ""                       ""                       ""
##                 A1Cresult                metformin              repaglinide
##                       ""                       ""                       ""
##              nateglinide           chlorpropamide              glimepiride
##                       ""                       ""                       ""
##                glipizide                glyburide              tolbutamide
##                       ""                       ""                       ""
##             pioglitazone            rosiglitazone                 acarbose
```

```
##                        ""                        ""                        ""
##                  miglitol                tolazamide                   insulin
##                        ""                        ""                        ""
##       glyburide.metformin       glipizide.metformin                    change
##                        ""                        ""                        ""
##               diabetesMed                readmitted
##                        ""                        ""
## PredictorMatrix:
##                          race gender age weight admission_type_id
## race                        0      1   1      1                 1
## gender                      1      0   1      1                 1
## age                         1      1   0      1                 1
## weight                      1      1   1      0                 1
## admission_type_id           1      1   1      1                 0
## discharge_disposition_id    1      1   1      1                 1
##                          discharge_disposition_id admission_source_id
## race                                            1                   1
## gender                                          1                   1
## age                                             1                   1
## weight                                          1                   1
## admission_type_id                               1                   1
## discharge_disposition_id                        0                   1
##                          time_in_hospital payer_code medical_specialty
## race                                    1          1                 1
## gender                                  1          1                 1
## age                                     1          1                 1
## weight                                  1          1                 1
## admission_type_id                       1          1                 1
## discharge_disposition_id                1          1                 1
##                          num_lab_procedures num_procedures num_medications
## race                                      1              1               1
## gender                                    1              1               1
## age                                       1              1               1
## weight                                    1              1               1
## admission_type_id                         1              1               1
## discharge_disposition_id                  1              1               1
##                          number_outpatient number_emergency number_inpatient
## race                                     1                1                1
## gender                                   1                1                1
## age                                      1                1                1
## weight                                   1                1                1
## admission_type_id                        1                1                1
## discharge_disposition_id                 1                1                1
##                          diag_1 diag_2 diag_3 number_diagnoses max_glu_serum
## race                          1      1      1                1             1
## gender                        1      1      1                1             1
## age                           1      1      1                1             1
## weight                        1      1      1                1             1
## admission_type_id             1      1      1                1             1
## discharge_disposition_id      1      1      1                1             1
##                          A1Cresult metformin repaglinide nateglinide
## race                             1         1           1           1
## gender                           1         1           1           1
## age                              1         1           1           1
```

```
## weight                        1         1           1           1
## admission_type_id             1         1           1           1
## discharge_disposition_id      1         1           1           1
##                          chlorpropamide glimepiride glipizide glyburide
## race                                  1           1         1         1
## gender                                1           1         1         1
## age                                   1           1         1         1
## weight                                1           1         1         1
## admission_type_id                     1           1         1         1
## discharge_disposition_id              1           1         1         1
##                          tolbutamide pioglitazone rosiglitazone acarbose
## race                               1            1             1        1
## gender                             1            1             1        1
## age                                1            1             1        1
## weight                             1            1             1        1
## admission_type_id                  1            1             1        1
## discharge_disposition_id           1            1             1        1
##                          miglitol tolazamide insulin glyburide.metformin
## race                            1          1       1                   1
## gender                          1          1       1                   1
## age                             1          1       1                   1
## weight                          1          1       1                   1
## admission_type_id               1          1       1                   1
## discharge_disposition_id        1          1       1                   1
##                          glipizide.metformin change diabetesMed readmitted
## race                                       1      1           1          1
## gender                                     1      1           1          1
## age                                        1      1           1          1
## weight                                     1      1           1          1
## admission_type_id                          1      1           1          1
## discharge_disposition_id                   1      1           1          1
## Number of logged events:  24
##   it im                      dep meth
## 1  1  1                   gender  pmm
## 2  1  1        admission_type_id  pmm
## 3  1  1 discharge_disposition_id  pmm
## 4  1  1      admission_source_id  pmm
## 5  1  2                   gender  pmm
## 6  1  2        admission_type_id  pmm
##
## 1                                     admission_type_id4, discharge_disposition_id10, discharge_dispo
## 2 discharge_disposition_id10, discharge_disposition_id12, discharge_disposition_id16, discharge_dispo
## 3
## 4
## 5                                     admission_type_id4, discharge_disposition_id10, discharge_dispo
## 6 discharge_disposition_id10, discharge_disposition_id12, discharge_disposition_id16, discharge_dispo
```

```r
completed_data <- complete(mice_diabetic_data,1)
```

A continuación se separará el dataset en 70% para entrenamiento y 30% para testing.

```r
# Split Data into Training and Testing in R
sample_size = floor(0.7*nrow(completed_data))
set.seed(777)
```

```r
# randomly split data in r
picked = sample(seq_len(nrow(completed_data)),size = sample_size)
development = completed_data[picked,]
holdout = completed_data[-picked,]
completed_data <- as.data.frame(completed_data)
development <- as.data.frame(development)
holdout <- as.data.frame(holdout)
```

Solución

Habiendo separado los datos, a continuación haremos tres métodos distintos para probar cuál ajusta mejor los datos. Los métodos utilizados son de aprendizaje supervisado de clasificación, para poder determinar la variable "readmitted".

Regresión logística multinomial

En el siguiente modelo, podemos observar que el ajuste es del 57.45% de los datos de prueba.

```r
#multinomial logistic regression

mylogit <- multinom(factor(readmitted) ~ ., data = development,MaxNWts =10000000)
```

```
## # weights:  249 (164 variable)
## initial  value 78260.744996
## iter  10 value 69706.149912
## iter  20 value 66982.766520
## iter  30 value 66001.180086
## iter  40 value 65688.502096
## iter  50 value 65456.073200
## iter  60 value 64353.815488
## iter  70 value 63490.088257
## iter  80 value 63014.956387
## iter  90 value 62651.346932
## iter 100 value 62597.171791
## final  value 62597.171791
## stopped after 100 iterations
```

```r
mylogit.results <- predict(mylogit,holdout)
results.logit <- data.frame(actual = holdout$readmitted, prediction = mylogit.results)
fit.logit <- count(filter(results.logit,actual==prediction))/count(results.logit) #0,5745
summary(mylogit)
```

```
## Warning in sqrt(diag(vc)): NaNs produced
```

```
## Call:
## multinom(formula = factor(readmitted) ~ ., data = development,
##     MaxNWts = 1e+07)
##
## Coefficients:
##   (Intercept)      race2      race3      race4       race5       race6
## 2   -3.072632  0.1241536 -0.3819508  0.1139894  0.003455878  0.01717105
## 3    1.586288 -0.2778335 -0.3289434 -0.2943949 -0.256544416 -0.06932425
##         gender2          age      weight admission_type_id2 admission_type_id3
```

```
## 2 -6.963512e-02  0.009697309  0.02030568         0.08678977         0.15951435
## 3  9.170349e-06 -0.038246986 -0.06275424        -0.07303811         0.09980852
##   admission_type_id4 admission_type_id7 discharge_disposition_id2
## 2         -0.7906000          -4.999170                -0.7329452
## 3         -0.3340478           2.857917                -0.5896725
##   discharge_disposition_id3 discharge_disposition_id4 discharge_disposition_id5
## 2                -0.4811332                -0.3102116                -1.0957595
## 3                -0.2854642                -0.2342167                -0.9523436
##   discharge_disposition_id6 discharge_disposition_id7 discharge_disposition_id8
## 2                -0.1301068                -0.4469961               -0.51533362
## 3                -0.2292192                -0.3913202                0.05284468
##   discharge_disposition_id9 discharge_disposition_id10
## 2                 -4.103243                   2.578413
## 3                 -1.507732                  -2.047938
##   discharge_disposition_id11 discharge_disposition_id12
## 2                  -1.371348                  -2.514847
## 3                   2.667130                  -2.048420
##   discharge_disposition_id13 discharge_disposition_id14
## 2                  -1.265464                  -2.325895
## 3                   1.289209                   1.188613
##   discharge_disposition_id15 discharge_disposition_id16
## 2                  -2.237289                  0.9361878
## 3                  -2.414008                  0.7363027
##   discharge_disposition_id17 discharge_disposition_id19
## 2                0.003784738                  -3.007266
## 3                0.292533184                   3.607987
##   discharge_disposition_id20 discharge_disposition_id22
## 2                 -0.8308331                  -1.478009
## 3                  1.0938636                  -1.169260
##   discharge_disposition_id23 discharge_disposition_id24
## 2                  0.2921033                 -0.2276716
## 3                  0.8094142                  0.1594510
##   discharge_disposition_id27 discharge_disposition_id28 admission_source_id2
## 2                  0.2077229                  -1.983627          -0.05430237
## 3                  0.7162271                  -1.598846           0.18152584
##   admission_source_id3 admission_source_id4 admission_source_id5
## 2           -0.1762249           -0.1064073            0.3835329
## 3           -0.4998636            0.4156230            0.7077654
##   admission_source_id6 admission_source_id7 admission_source_id8
## 2           -0.3270872           0.16630098           -0.1892185
## 3            0.2390790          -0.03301481           -0.1442567
##   admission_source_id10 admission_source_id11 admission_source_id13
## 2              0.630911             -1.099233            -0.6816109
## 3              1.012247              1.426942             0.8935965
##   admission_source_id14 admission_source_id22 admission_source_id25
## 2                     0            -0.7832817            -0.7373866
## 3                     0            -0.2302626             0.9870382
##   time_in_hospital  payer_code medical_specialty num_lab_procedures
## 2      0.001196352 0.006396545      -0.001438021       0.0002065351
## 3     -0.015684038 0.010852436       0.001869930      -0.0007427257
##   num_procedures num_medications number_outpatient number_emergency
## 2    -0.01854032    -0.002863530        0.03801892     -0.009000476
## 3     0.01531607    -0.006132582       -0.05682655     -0.214491220
##   number_inpatient       diag_1       diag_2       diag_3 number_diagnoses
```

```
## 2       -0.1310246 0.0002009613 0.0000867820 -3.961618e-05      0.009740744
## 3       -0.4926653 0.0003731535 0.0002337968 -8.416459e-05     -0.058906976
##   max_glu_serum   A1Cresult  metformin repaglinide nateglinide chlorpropamide
## 2   -0.02614619 -0.03930791 0.04292613  0.01124840  0.20904380      0.7398067
## 3    0.03556358 -0.02449675 0.16886741 -0.08378053  0.05770189      0.9527435
##   glimepiride  glipizide   glyburide tolbutamide pioglitazone rosiglitazone
## 2 -0.04759913 0.05481124 -0.02852433 -0.04052023   0.14397780    0.13653372
## 3 -0.02473244 0.00795353  0.02079189  0.40348006   0.06115466    0.06662578
##     acarbose  miglitol  tolazamide     insulin glyburide.metformin
## 2 0.51320827 0.5612213 -0.08909419 0.003414465          0.09159215
## 3 0.07562085 0.2740847  0.17498787 0.035168807         -0.28597001
##   glipizide.metformin      change diabetesMed
## 2          -0.3579456 0.002233954 -0.07479616
## 3          -1.2632603 0.007842423 -0.33934810
##
## Std. Errors:
##   (Intercept)      race2      race3      race4      race5      race6    gender2
## 2 0.007632529 0.03904198 0.03491920 0.03623140 0.03996840 0.03584972 0.02646273
## 3 0.007738862 0.03858884 0.03862303 0.03543472 0.04076853 0.03770491 0.02588905
##           age     weight admission_type_id2 admission_type_id3
## 2 0.009214153 0.01090220         0.03706140         0.04406228
## 3 0.009000853 0.01123788         0.03678191         0.04296474
##   admission_type_id4 admission_type_id7 discharge_disposition_id2
## 2       0.0001167342       6.318381e-08                0.03185058
## 3       0.0001565358       1.048541e-05                0.03334589
##   discharge_disposition_id3 discharge_disposition_id4 discharge_disposition_id5
## 2                0.03724588                0.04653337                0.03900982
## 3                0.03645874                0.04818207                0.04411405
##   discharge_disposition_id6 discharge_disposition_id7 discharge_disposition_id8
## 2                0.03773718                0.05433390              0.0003443325
## 3                0.03791585                0.05462276              0.0004085044
##   discharge_disposition_id9 discharge_disposition_id10
## 2              1.522724e-05               9.404612e-06
## 3              1.202403e-04               1.924682e-06
##   discharge_disposition_id11 discharge_disposition_id12
## 2               0.0004449868               2.188103e-05
## 3               0.0011425400               6.121444e-05
##   discharge_disposition_id13 discharge_disposition_id14
## 2               0.0004107001               0.0001490949
## 3               0.0009709889               0.0009455949
##   discharge_disposition_id15 discharge_disposition_id16
## 2               0.0004651733               2.343101e-05
## 3               0.0004696310               2.261526e-05
##   discharge_disposition_id17 discharge_disposition_id19
## 2               5.618003e-05               2.450008e-07
## 3               6.107200e-05               3.989951e-06
##   discharge_disposition_id20 discharge_disposition_id22
## 2               2.076322e-06                 0.04905329
## 3               7.104727e-06                 0.04789285
##   discharge_disposition_id23 discharge_disposition_id24
## 2                0.002072667               0.0001833571
## 3                0.002183319               0.0001992317
##   discharge_disposition_id27 discharge_disposition_id28 admission_source_id2
## 2               7.844181e-05               0.0003925323           0.03892054
```

```
## 3               8.252062e-05              0.0005811017          0.04196454
##   admission_source_id3 admission_source_id4 admission_source_id5
## 2          0.001348172           0.02550786           0.04780846
## 3          0.001441243           0.02768140           0.05010112
##   admission_source_id6 admission_source_id7 admission_source_id8
## 2           0.04027615           0.03578854         0.0001984316
## 3           0.04114213           0.03528265         0.0002561030
##   admission_source_id10 admission_source_id11 admission_source_id13
## 2          8.251293e-05          2.434958e-06          2.805504e-06
## 3          8.851531e-05          5.897226e-06          5.629659e-06
##   admission_source_id14 admission_source_id22 admission_source_id25
## 2                   NaN          9.241949e-05          7.802113e-06
## 3                   NaN          1.127217e-04          1.913863e-05
##   time_in_hospital  payer_code medical_specialty num_lab_procedures
## 2      0.005089274 0.002877669      0.0008226443       0.0007507754
## 3      0.005007883 0.002805824      0.0007917715       0.0007352029
##   num_procedures num_medications number_outpatient number_emergency
## 2    0.008919972     0.002088870       0.009657262      0.009602157
## 3    0.008644083     0.002032074       0.010763530      0.016614658
##   number_inpatient       diag_1       diag_2       diag_3 number_diagnoses
## 2      0.007972875 8.217012e-05 8.787815e-05 7.511053e-05      0.008071991
## 3      0.010335040 8.040532e-05 8.563908e-05 7.331442e-05      0.007798585
##   max_glu_serum  A1Cresult  metformin repaglinide nateglinide chlorpropamide
## 2    0.03948058 0.02567947 0.03248961  0.03504224  0.04573264     0.01771898
## 3    0.03900126 0.02505740 0.03154265  0.03461144  0.04525283     0.01802898
##   glimepiride  glipizide  glyburide tolbutamide pioglitazone rosiglitazone
## 2  0.04975210 0.03552097 0.03766180 0.007880171   0.04587704    0.04801286
## 3  0.04865377 0.03499660 0.03652582 0.008014223   0.04529966    0.04733137
##     acarbose   miglitol  tolazamide    insulin glyburide.metformin
## 2 0.03593826 0.01626723 0.008015768 0.01533598          0.04736183
## 3 0.03533832 0.01647184 0.008162769 0.01525124          0.04696981
##   glipizide.metformin     change diabetesMed
## 2         0.007749613 0.02950072  0.03938185
## 3         0.007842860 0.02906224  0.03824198
##
## Residual Deviance: 125194.3
## AIC: 125518.3
```

SVM

En el siguiente modelo, podemos observar que el ajuste es del 57.48% de los datos de prueba.

```r
# svm supervised
svm.model <- e1071::svm(readmitted ~ ., data = as.data.frame(development))
svm.results <- predict(svm.model,holdout)
results.svm <- data.frame(actual = na.omit(holdout)$readmitted, prediction = svm.results)
fit.svm <- count(filter(results.svm,actual==round(prediction)))/count(results.svm) # 0.5748772
```

K Nearest Neighbors

Por último, veremos el ajuste al modelo k nearest neighbors, obteniendo un ajuste del 52.18% en los datos de prueba:

```r
#knn
knn.model <- knn(na.omit(development),na.omit(holdout),factor(na.omit(development)$readmitted),k=20)
tab <- table(knn.model,as.data.frame(na.omit(holdout))$readmitted)
results.knn <- data.frame(actual = as.data.frame(na.omit(holdout))$readmitted,prediction = knn.model)
fit.knn <- count(filter(results.knn,actual==prediction))/count(results.knn) # 0.5218146
```

Conclusiones

Comparando los tres modelos, podemos ver que el que tiene mayor accuracy es Support Vector Machine, con un porcentaje de acierto del 57.48%, seguido de la regresión logística multinomial, con un porcentaje de acierto del 57.45%, y por último K nearest neighbors con un porcentaje del 52.18%.