# Clasificador diseasescodes TCGA sondas con mayor variabilidad

Alberto Joven Álvarez

28 de octubre, 2022

## Índice



```
library(knitr)
library(SummarizedExperiment)
library(GEOquery)
library(Rtsne)
library(ggplot2)
library(dplyr)
library(sva)
library(caret)
library(class)
library(C50)
library(gmodels)
library(keras)
library(randomForest)
library(googledrive)
library(doParallel)
```

```
library(TCGAutils)
registerDoParallel(cores=3)
```

# 1 Clasificador machine learning de muestras tumorales.

## 1.1 Descripción datos de entrada

En este clasificador, como datos de entrada, se utilizaránn las sondas seleccionadas con el criterio de selección d elas 10.000 sondas con mayor variabilidad. Arrays: Illumina methylation 450k.

## 1.2 Descarga de las anotaciones adicionales de las sondas

```
elist <- getGEO("GSE42409")
GSE42409 <- elist[[1]] %>% featureData()

# chequeo variables GSE42409
sum(GSE42409$`Target CpG SNP` != "")
```

```
## [1] 17098
```

```
table(GSE42409$`n_target CpG SNP`, useNA = "always")
```

```
##
##      1      2      3      4   <NA>
##  16712    376      9      1 411118
```

```
sum(GSE42409$SNPprobe != "")
```

```
## [1] 107692
```

```
table(GSE42409$n_SNPprobe, useNA="always")
```

```
##
##      1      2      3      4      5      6      7      8      9     10     11
##  87303  16199   2972    656    217    113     61     38     26     10      5
##     12     13     14     15     16     17     20     23     24     26     27
##     10      7      6      6      6      2      2      1      1      1      5
##     30     31     33     37     38     40     41     44     46     47     48
##      3      2      2      2      1      1      2      1      1      2      3
##     49     50     51   <NA>
##      1     22      2 320524
```

```
table(GSE42409$n_bp_repetitive, useNA="always")
```

```
##
##      0      1      2      3      4      5      6      7      8      9     10
## 355423    760    945   1108   1464    977    937   1202   1218   1047    953
##     11     12     13     14     15     16     17     18     19     20     21
##    886    780    733    701    723    709    700    674    764    748    731
##     22     23     24     25     26     27     28     29     30     31     32
##    686    644    669    636    634    600    637    623    638    554    563
##     33     34     35     36     37     38     39     40     41     42     43
##    624    590    551    591    593    609    592    588    602    599    614
##     44     45     46     47     48     49     50   <NA>
##    585    640    733   1355   3746   9717  23820      0
```

```
table(GSE42409$AlleleA_Hits, useNA = "always")
```

```
##
##      1      2      3      4      5      6      7      8      9     31   <NA>
## 427745    378     57     13      8      6      6      1      1      1      0
```

```
table(GSE42409$AlleleB_Hits, useNA = "always")
```

```
##
##      0      1   <NA>
## 312164 116052      0
```

## 1.3 Carga inicial de los datos

Carga del objeto **SummarizedExperiment** que contiene los valores betas, valores de fenotipos y rangos de todas las sondas y todos los pacientes de todos 33 estudios existentes en TCGA referidos a metilación, analizados con el array Illumina 450k.

Únicamente se han excluido las sondas con NAs en más de un 10% de las observaciones.

```
################################################################
# ESTE SCRIPT NO SE EJECUTA EN EL CUADERNO RMARKDOWN PORQUE DA
# ERROR DE MEMORIA SOBREPASADA. EL ARCHIVO
# SummarizedExperimen.Rda tiene un tamaño de más de 30 Gb.
# SE HA EJECUTADO COMO SCRIPT INDEPENDIENTE
# Y PARA EL ANÁLISIS SE CARGAN LOS FICHEROS FINALES RESULTANTES
################################################################


# Carga del objeto Summarized Experiment con 9707 muestras
# y 395312 sondas
# se han eliminado previamente las sondas con más del 10% de muestras con NAs
# sondas del array completo 485577

load("G:/TFM UOC/datos/SummarizedExperiment.Rda")
data

nombres_sondas <- row.names(data)
substring(nombres_sondas, 1 , 2) %>% table()
```

```r
# sondas a excluir:
# Las que comienzan con rs o ch 1.674 sondas
sondas_excluir_1 <- nombres_sondas[substring(nombres_sondas, 1 , 2 ) %in% c("ch", "rs")]

# Las que tienen en la anotación adicional el campo `Target CpG SNP`no vacío
sondas_excluir_2 <- nombres_sondas[sondas_en_GSE$`Target CpG SNP` != ""]

# Las que tiene más de una localización in silico:
sondas_excluir_3 <- nombres_sondas[sondas_en_GSE$AlleleA_Hits != 1]
sondas_excluir_4 <- nombres_sondas[sondas_en_GSE$AlleleB_Hits != 0]

# Las que apuntan a  ADN repetitivo
sondas_excluir_5 <- nombres_sondas[sondas_en_GSE$n_bp_repetitive != 0]


s1 <- union(sondas_excluir_1, sondas_excluir_2)
s1 <- union(sondas_excluir_3, s1)
s1 <- union(sondas_excluir_4, s1)
s1 <- union(sondas_excluir_5, s1)

data_sondas_dep <- data[!(nombres_sondas %in% s1)]
data_sondas_dep
rm(data)

save(data_sondas_dep, file="G:/TFM UOC/datos/Clasificador_variabilidad/data_sondas_dep.Rda")

#########################################################
# desglose entre muestras de entrenamiento y de test
#########################################################

set.seed(123)
etiqueta <- data_sondas_dep$label

in_train <- createDataPartition(etiqueta, p=0.75, list=FALSE)

train <- data_sondas_dep[ , in_train]
test <- data_sondas_dep[ , -in_train]

save(train, file="G:/TFM UOC/datos/Clasificador_variabilidad/data_sondas_dep_train.Rda")
save(test, file="G:TFM UOC/datos/Clasificador_variabilidad/data_sondas_dep_test.Rda")

rm(data_sondas_dep)

#########################################################################
# selección de las 10000 sondas con mayor variabilidad del grupo train
#########################################################################

betas_train <- assay(train, "counts")
sds <- apply(betas_train, 1, sd)
orden <- order(sds, decreasing=T)
sds_o <- sds[orden][1:10000]

summarized_train <- train[names(sds_o) ]
```

```
summarized_test <- test[names(sds_o) ]


save(summarized_train, file="G:/TFM UOC/datos/Clasificador_variabilidad/summarized_train.Rda")
save(summarized_test, file="G:TFM UOC/datos/Clasificador_variabilidad/summarized_test.Rda")


rm(train, test, betas_train)
```

### 1.4 Carga de los ficheros train y test

```
load("G:/TFM UOC/datos/Clasificador_variabilidad/summarized_train.Rda")
load("G:/TFM UOC/datos/Clasificador_variabilidad/summarized_test.Rda")

train <- summarized_train
test <- summarized_test
```

### 1.5 Tabla de distribución de muestras entre train y test

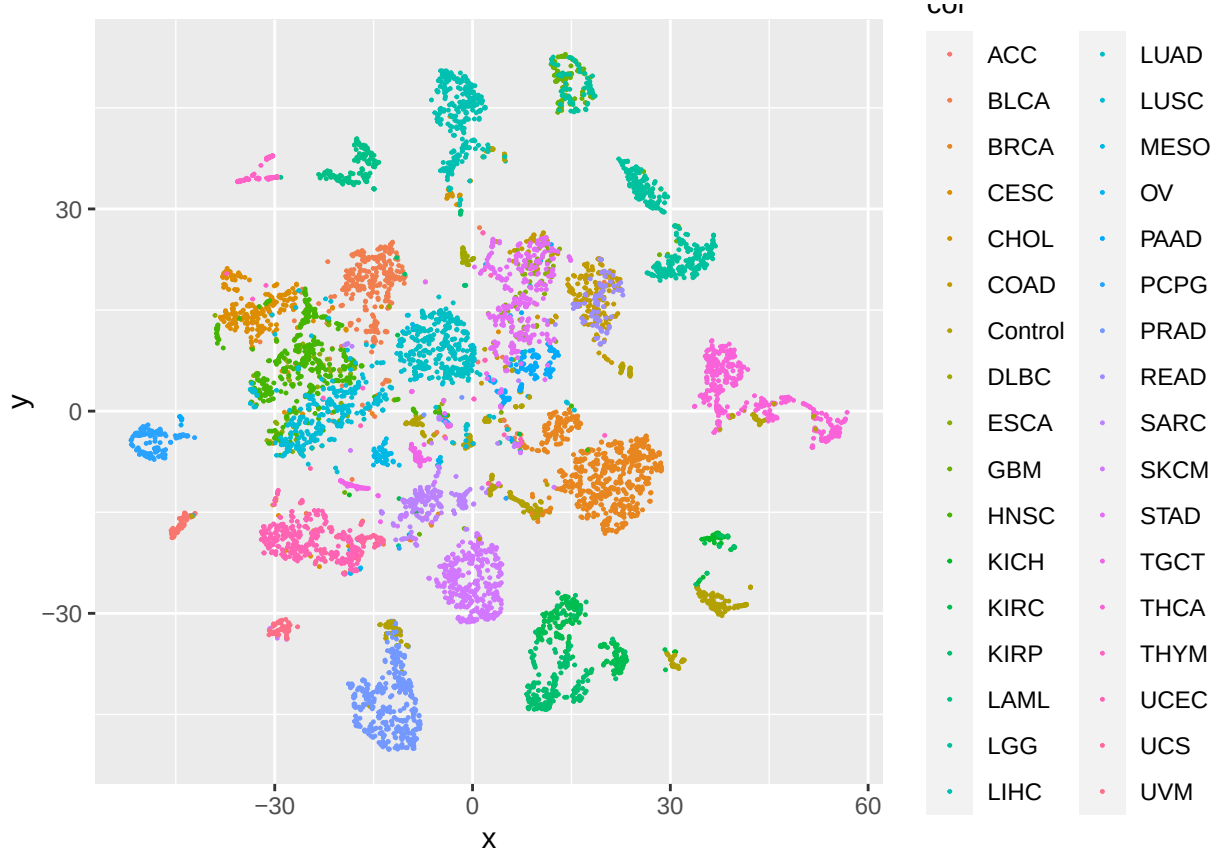
```
t1 <- train$label %>% table() %>% as.matrix()
t2 <- test$label %>% table() %>% as.matrix()

df <- data.frame(Train= table(train$label),
                 Test = table(test$label),
                 Total = t1+t2)

df[, c(2,4,5) ] %>% kable()
```

| | Train.Freq | Test.Freq | Total |
|---|---|---|---|
| ACC | 60 | 20 | 80 |
| BLCA | 310 | 103 | 413 |
| BRCA | 591 | 197 | 788 |
| CESC | 232 | 77 | 309 |
| CHOL | 27 | 9 | 36 |
| COAD | 222 | 73 | 295 |
| Control | 551 | 183 | 734 |
| DLBC | 36 | 12 | 48 |
| ESCA | 140 | 46 | 186 |
| GBM | 115 | 38 | 153 |
| HNSC | 398 | 132 | 530 |
| KICH | 50 | 16 | 66 |
| KIRC | 240 | 80 | 320 |
| KIRP | 207 | 69 | 276 |
| LAML | 146 | 48 | 194 |
| LGG | 398 | 132 | 530 |
| LIHC | 285 | 94 | 379 |
| LUAD | 345 | 115 | 460 |
| LUSC | 278 | 92 | 370 |

| | Train.Freq | Test.Freq | Total |
|---|---|---|---|
| MESO | 66 | 21 | 87 |
| OV | 8 | 2 | 10 |
| PAAD | 139 | 46 | 185 |
| PCPG | 138 | 46 | 184 |
| PRAD | 375 | 124 | 499 |
| READ | 75 | 24 | 99 |
| SARC | 199 | 66 | 265 |
| SKCM | 355 | 118 | 473 |
| STAD | 297 | 98 | 395 |
| TGCT | 105 | 34 | 139 |
| THCA | 384 | 127 | 511 |
| THYM | 93 | 31 | 124 |
| UCEC | 324 | 108 | 432 |
| UCS | 43 | 14 | 57 |
| UVM | 60 | 20 | 80 |

### 1.6 Gráfico previo Rtsne

```
betas_train <- assay(summarized_train, "counts") %>% t()
etiqueta <- colData(summarized_train)$label

sed.seed=123
tsne <- Rtsne(betas_train, partial_pca=TRUE, dims=2, perplexity=30, verbose =FALSE, max_iter=1000 )

# Gráfico por patologías
tsne_plot <- data.frame(x = tsne$Y[,1], y = tsne$Y[,2], col = etiqueta)
ggplot(tsne_plot) + geom_point(aes(x=x, y=y, color=col), size=0.2)
```




## 1.7 Tabla ubicaciones sondas seleccionadas

```
rowRanges(train) %>% seqnames() %>% table()
```

```
## .
##  chr1  chr2  chr3  chr4  chr5  chr6  chr7  chr8  chr9 chr10 chr11 chr12 chr13
##   898   811   519   397   585   717   847   563   124   527   456   528   293
## chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22  chrX  chrY
##   286   241   379   508   129   412   209    69   100   399     3
```

## 1.8 Tabla con la tipología de las sondas seleccionadas

```
rowRanges(train)$channel %>% table() %>% kable()
```

| . | Freq |
|---|---|
| Both | 4602 |
| Grn | 818 |
| Red | 4580 |

## 1.9 Tabla con tipos de targets HIL en las sondas seleccionadas

```
GSE42409$HIL_CpG_class[GSE42409$ID %in% names(train)]  %>%  table() %>%  kable()
```

| . | Freq |
|---|---|
| HC | 2737 |
| IC | 2740 |
| ICshore | 944 |
| LC | 2870 |

## 1.10 Estimación efecto batch de la variable plate_id

**ESTE CÓDIGO NO SE EJECUTA**

```
edata <- assay(train, "counts")
pdata <- colData(train)

nombres <- assay(train, "counts") %>% colnames()
plate_id <- TCGAbiospec(nombres)
plate_id <- plate_id$plate

train$plate_id <- plate_id

mod0 <- model.matrix( ~ 1, data = pdata)
mod <- model.matrix( ~ as.factor(label), data = pdata)

combat_edata <- ComBat(dat = edata, batch = plate_id,
                       mod = mod0, par.prior=TRUE)

assay(summarized_train, "counts") <- combat_edata
```

## 1.11 Algoritmo red neuronal

### 1.11.1 Preparación datos

```
fenotipos_train <- colData(summarized_train)$label %>% factor(ordered=TRUE)
fenotipos_test <- colData(summarized_test)$label %>% factor(ordered=TRUE)
betas_train <- assay(summarized_train, "counts") %>% t()
betas_test <- assay(summarized_test, "counts") %>% t()
```

### 1.11.2 Formulación del modelo

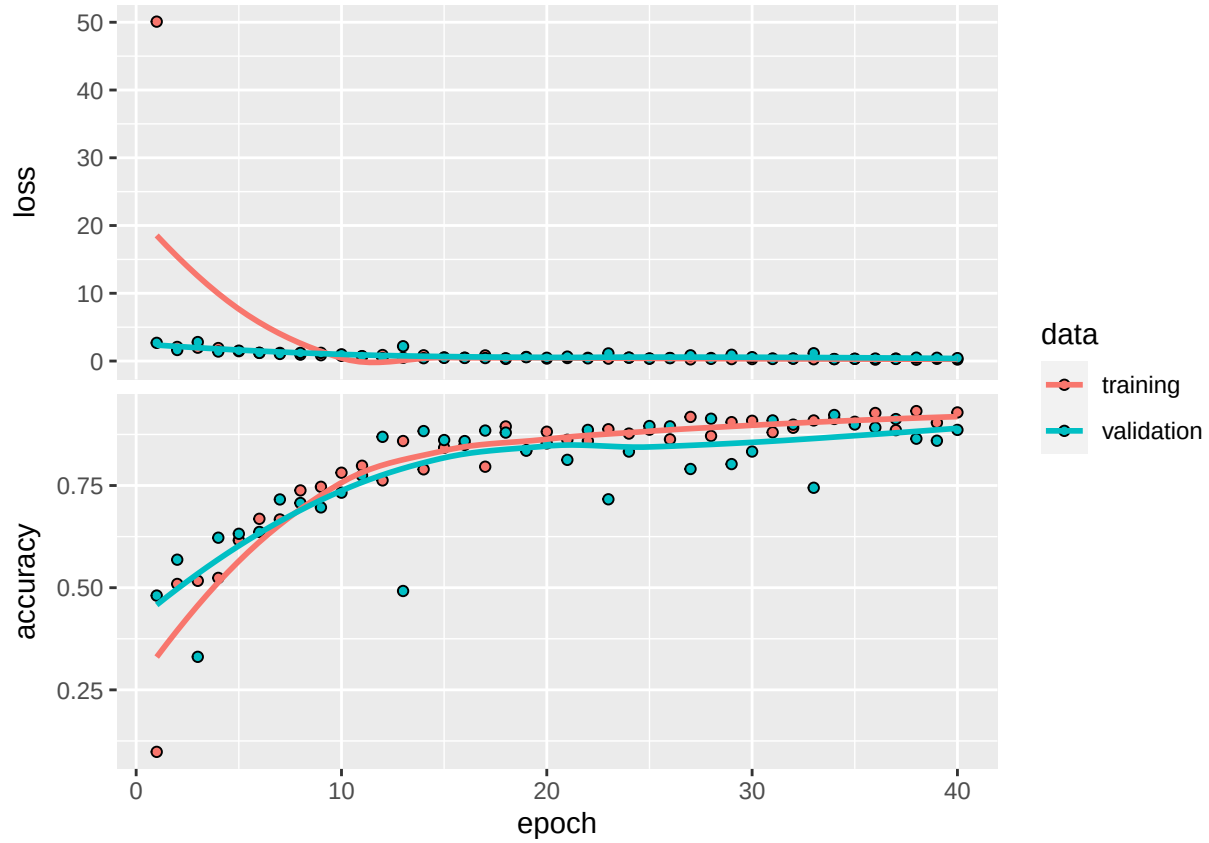
```
red <- keras_model_sequential() %>%
  layer_dense(units = 1000, activation="relu", input_shape=10000) %>%
  layer_dense(units = 200) %>%
  layer_dense(units = 35, activation ="softmax")
```

```
red %>% compile(
  optimizer = "rmsprop",
  loss = "categorical_crossentropy",
  metrics = c("accuracy")
)

train_labels <- to_categorical(as.integer(fenotipos_train))
test_labels <- to_categorical(as.integer(fenotipos_test))

red
```

```
## Model: "sequential"
## _________________________________________________________________________
##  Layer (type)                       Output Shape                    Param #
## =========================================================================
##  dense_2 (Dense)                    (None, 1000)                    10001000
##  dense_1 (Dense)                    (None, 200)                     200200
##  dense (Dense)                      (None, 35)                      7035
## =========================================================================
## Total params: 10,208,235
## Trainable params: 10,208,235
## Non-trainable params: 0
## _________________________________________________________________________
```

### 1.11.3 Entrenamiento del modelo

```
hist <- red %>% fit(betas_train, train_labels,
                    epoch=40, batch_size=512,
                    validation_data = list(betas_test, test_labels))
plot(hist)
```




### 1.11.4 Evaluación del modelo

```
metrics <- red %>% evaluate(betas_test, test_labels)
metrics
```

```
##      loss  accuracy
## 0.4319769 0.8861284
```

### 1.11.5 Matriz de confusión con el subset test

```
prediccion <- red %>% predict(betas_test) %>% k_argmax() %>%
              as.array() %>% as.integer()

l <- as.list(1:34)
names(l) <- levels(fenotipos_test)
f <- names(l)[prediccion]
lev <- levels(fenotipos_test)
prediccion <- factor(f, levels=lev)

c3 <- confusionMatrix(fenotipos_test, prediccion)
c3
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction ACC BLCA BRCA CESC CHOL COAD Control DLBC ESCA GBM HNSC KICH KIRC
##    ACC      19    0    0    0    0    0       0    0    0   0    0    0    0
##    BLCA      0   87    0    0    0    0       2    1    1   0    0    0    0
##    BRCA      0    0  194    0    0    0       1    0    0   0    0    0    0
##    CESC      0    0    0   72    0    0       1    0    0   0    1    0    0
##    CHOL      0    0    0    0    6    0       1    0    0   0    0    0    0
##    COAD      0    0    0    0    0   72       0    0    0   0    0    0    0
##    Control   0    0    0    0    0    0     151    0    0   0    0    0    0
##    DLBC      0    0    0    0    0    0       0   11    0   0    0    0    0
##    ESCA      0    0    0    0    0    0       1    0   40   0    0    0    0
##    GBM       0    0    0    0    0    0       0    0    0  19    0    0    0
##    HNSC      0    0    0    0    0    0       1    0   34   0   72    0    0
##    KICH      0    0    0    0    0    0       0    0    0   0    0   15    0
##    KIRC      0    0    0    0    0    0       0    0    0   0    0    1   62
##    KIRP      0    1    0    0    0    0       0    0    0   0    0    2    1
##    LAML      0    0    0    0    0    0       0    0    0   0    0    0    0
##    LGG       0    0    0    0    0    0       0    0    0   5    0    0    0
##    LIHC      0    0    0    0    0    0       0    0    0   0    0    0    0
##    LUAD      0    0    0    0    0    0       0    0    0   0    0    0    0
##    LUSC      0    0    0    0    0    0       0    0    1   0    0    0    0
##    MESO      0    0    0    0    0    0       0    0    0   0    0    0    0
##    OV        0    0    0    0    0    0       0    0    0   0    0    0    0
##    PAAD      0    0    0    0    0    0       2    0    0   0    0    0    0
##    PCPG      0    0    0    0    0    0       1    0    0   0    0    1    0
##    PRAD      0    0    0    0    0    0       3    0    0   0    0    0    0
##    READ      0    0    0    0    0   24       0    0    0   0    0    0    0
##    SARC      0    0    0    0    0    0       0    0    0   0    0    0    0
##    SKCM      0    0    0    0    0    0       1    0    0   0    0    0    0
##    STAD      0    0    0    0    0    2       2    1   24   0    0    0    0
##    TGCT      0    0    0    0    0    0       0    0    0   0    0    0    0
##    THCA      0    0    0    0    0    0       0    0    0   0    0    0    0
##    THYM      0    0    0    0    0    0       0    0    0   0    0    0    0
##    UCEC      0    0    0    0    0    0       0    0    0   0    0    0    0
##    UCS       0    0    0    0    0    0       0    0    0   0    0    0    0
##    UVM       0    0    0    0    0    0       0    0    0   0    0    0    0
##            Reference
## Prediction KIRP LAML LGG LIHC LUAD LUSC MESO  OV PAAD PCPG PRAD READ SARC SKCM
##    ACC        0    0   0    0    1    0    0   0    0    0    0    0    0    0
##    BLCA       5    0   0    0    0    4    0   0    0    0    0    0    0    0
##    BRCA       0    0   0    0    0    0    0   0    0    0    0    0    2    0
##    CESC       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##    CHOL       0    0   0    2    0    0    0   0    0    0    0    0    0    0
##    COAD       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##    Control    0    0   0    1    0    0    0   0    0    0    2    0    0    0
##    DLBC       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##    ESCA       0    0   0    0    0    2    0   0    0    0    0    0    0    0
##    GBM        0    0  19    0    0    0    0   0    0    0    0    0    0    0
##    HNSC       0    0   0    0    0   25    0   0    0    0    0    0    0    0
##    KICH       1    0   0    0    0    0    0   0    0    0    0    0    0    0
##    KIRC      11    0   0    0    0    0    0   0    0    0    0    0    2    0
##    KIRP      65    0   0    0    0    0    0   0    0    0    0    0    0    0
```

```
##     LAML        0   48   0    0    0    0    0  0    0    0    0    0    0    0
##     LGG         0    0 126    0    0    0    0  0    0    0    0    0    0    0
##     LIHC        0    0   0   93    0    0    0  0    0    0    0    0    0    0
##     LUAD        0    0   0    1  105    5    0  0    0    0    0    0    0    0
##     LUSC        0    0   0    0    5   82    0  0    0    0    0    0    0    0
##     MESO        0    0   0    0    0    0   21  0    0    0    0    0    0    0
##     OV          0    0   0    0    0    0    0  1    0    0    0    0    0    0
##     PAAD        0    0   0    0    0    0    0  0   42    0    0    0    0    0
##     PCPG        0    0   0    0    0    0    0  0    0   44    0    0    0    0
##     PRAD        0    0   0    0    0    0    0  0    0    0  121    0    0    0
##     READ        0    0   0    0    0    0    0  0    0    0    0    0    0    0
##     SARC        0    0   0    0    0    0    0  0    0    0    0    0   63    1
##     SKCM        0    0   0    0    0    0    0  0    0    0    0    0    0  116
##     STAD        0    0   0    3    1    0    0  0    1    0    0    0    0    0
##     TGCT        0    0   0    0    0    0    0  0    0    0    0    0    0    0
##     THCA        0    0   0    0    0    0    0  0    0    0    0    0    0    0
##     THYM        0    0   0    0    0    0    0  0    0    0    0    0    0    0
##     UCEC        0    0   0    0    0    0    0  0    0    0    0    0    0    0
##     UCS         0    0   0    0    0    0    0  0    0    0    0    0    0    0
##     UVM         0    0   0    0    0    0    0  0    0    0    0    0    0    0
##              Reference
## Prediction STAD TGCT THCA THYM UCEC UCS UVM
##     ACC       0    0    0    0    0   0   0
##     BLCA      0    0    3    0    0   0   0
##     BRCA      0    0    0    0    0   0   0
##     CESC      0    0    0    0    3   0   0
##     CHOL      0    0    0    0    0   0   0
##     COAD      1    0    0    0    0   0   0
##     Control   0    0   28    1    0   0   0
##     DLBC      0    0    1    0    0   0   0
##     ESCA      3    0    0    0    0   0   0
##     GBM       0    0    0    0    0   0   0
##     HNSC      0    0    0    0    0   0   0
##     KICH      0    0    0    0    0   0   0
##     KIRC      0    0    4    0    0   0   0
##     KIRP      0    0    0    0    0   0   0
##     LAML      0    0    0    0    0   0   0
##     LGG       0    0    1    0    0   0   0
##     LIHC      0    0    1    0    0   0   0
##     LUAD      0    0    4    0    0   0   0
##     LUSC      0    0    4    0    0   0   0
##     MESO      0    0    0    0    0   0   0
##     OV        0    0    0    0    1   0   0
##     PAAD      0    0    2    0    0   0   0
##     PCPG      0    0    0    0    0   0   0
##     PRAD      0    0    0    0    0   0   0
##     READ      0    0    0    0    0   0   0
##     SARC      0    0    2    0    0   0   0
##     SKCM      0    0    1    0    0   0   0
##     STAD     64    0    0    0    0   0   0
##     TGCT      0   34    0    0    0   0   0
##     THCA      0    0  127    0    0   0   0
##     THYM      0    0    0   31    0   0   0
##     UCEC      0    0    0    0  107   1   0
```

```
##    UCS        0    0    0    0    4  10   0
##    UVM        0    0    0    0    0   0  20
##
## Overall Statistics
##
##                Accuracy : 0.8861
##                  95% CI : (0.8728, 0.8985)
##     No Information Rate : 0.0803
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8809
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL
## Sensitivity            1.000000     0.98864     1.00000     1.00000    1.000000
## Specificity            0.999583     0.99312     0.99865     0.99787    0.998755
## Pos Pred Value         0.950000     0.84466     0.98477     0.93506    0.666667
## Neg Pred Value         1.000000     0.99957     1.00000     1.00000    1.000000
## Prevalence             0.007867     0.03644     0.08033     0.02981    0.002484
## Detection Rate         0.007867     0.03602     0.08033     0.02981    0.002484
## Detection Prevalence   0.008282     0.04265     0.08157     0.03188    0.003727
## Balanced Accuracy      0.999791     0.99088     0.99932     0.99893    0.999377
##                      Class: COAD Class: Control Class: DLBC Class: ESCA
## Sensitivity              0.73469         0.90419    0.846154     0.40000
## Specificity              0.99957         0.98577    0.999584     0.99741
## Pos Pred Value           0.98630         0.82514    0.916667     0.86957
## Neg Pred Value           0.98890         0.99283    0.999168     0.97467
## Prevalence               0.04058         0.06915    0.005383     0.04141
## Detection Rate           0.02981         0.06253    0.004555     0.01656
## Detection Prevalence     0.03023         0.07578    0.004969     0.01905
## Balanced Accuracy        0.86713         0.94498    0.922869     0.69870
##                      Class: GBM Class: HNSC Class: KICH Class: KIRC Class: KIRP
## Sensitivity            0.791667     0.98630    0.789474     0.98413     0.79268
## Specificity            0.992054     0.97438    0.999583     0.99235     0.99829
## Pos Pred Value         0.500000     0.54545    0.937500     0.77500     0.94203
## Neg Pred Value         0.997897     0.99956    0.998333     0.99957     0.99275
## Prevalence             0.009938     0.03023    0.007867     0.02609     0.03395
## Detection Rate         0.007867     0.02981    0.006211     0.02567     0.02692
## Detection Prevalence   0.015735     0.05466    0.006625     0.03313     0.02857
## Balanced Accuracy      0.891860     0.98034    0.894528     0.98824     0.89548
##                      Class: LAML Class: LGG Class: LIHC Class: LUAD Class: LUSC
## Sensitivity              1.00000    0.86897     0.93000     0.93750     0.69492
## Specificity              1.00000    0.99736     0.99957     0.99566     0.99565
## Pos Pred Value           1.00000    0.95455     0.98936     0.91304     0.89130
## Neg Pred Value           1.00000    0.99168     0.99698     0.99696     0.98450
## Prevalence               0.01988    0.06004     0.04141     0.04638     0.04886
## Detection Rate           0.01988    0.05217     0.03851     0.04348     0.03395
## Detection Prevalence     0.01988    0.05466     0.03892     0.04762     0.03810
## Balanced Accuracy        1.00000    0.93316     0.96478     0.96658     0.84528
##                      Class: MESO Class: OV Class: PAAD Class: PCPG Class: PRAD
## Sensitivity             1.000000 1.0000000     0.97674     1.00000     0.98374
```

```
## Specificity             1.000000 0.9995857     0.99831     0.99916     0.99869
## Pos Pred Value          1.000000 0.5000000     0.91304     0.95652     0.97581
## Neg Pred Value          1.000000 1.0000000     0.99958     1.00000     0.99913
## Prevalence              0.008696 0.0004141     0.01781     0.01822     0.05093
## Detection Rate          0.008696 0.0004141     0.01739     0.01822     0.05010
## Detection Prevalence    0.008696 0.0008282     0.01905     0.01905     0.05135
## Balanced Accuracy       1.000000 0.9997929     0.98753     0.99958     0.99122
##                      Class: READ Class: SARC Class: SKCM Class: STAD
## Sensitivity                   NA     0.94030     0.99145     0.94118
## Specificity             0.990062     0.99872     0.99913     0.98551
## Pos Pred Value                NA     0.95455     0.98305     0.65306
## Neg Pred Value                NA     0.99830     0.99956     0.99827
## Prevalence              0.000000     0.02774     0.04845     0.02816
## Detection Rate          0.000000     0.02609     0.04803     0.02650
## Detection Prevalence    0.009938     0.02733     0.04886     0.04058
## Balanced Accuracy             NA     0.96951     0.99529     0.96334
##                      Class: TGCT Class: THCA Class: THYM Class: UCEC Class: UCS
## Sensitivity              1.00000     0.71348     0.96875     0.93043   0.909091
## Specificity              1.00000     1.00000     1.00000     0.99957   0.998336
## Pos Pred Value           1.00000     1.00000     1.00000     0.99074   0.714286
## Neg Pred Value           1.00000     0.97771     0.99958     0.99653   0.999584
## Prevalence               0.01408     0.07371     0.01325     0.04762   0.004555
## Detection Rate           0.01408     0.05259     0.01284     0.04431   0.004141
## Detection Prevalence     0.01408     0.05259     0.01284     0.04472   0.005797
## Balanced Accuracy        1.00000     0.85674     0.98438     0.96500   0.953714
##                      Class: UVM
## Sensitivity            1.000000
## Specificity            1.000000
## Pos Pred Value         1.000000
## Neg Pred Value         1.000000
## Prevalence             0.008282
## Detection Rate         0.008282
## Detection Prevalence   0.008282
## Balanced Accuracy      1.000000
```

## 1.12 Algoritmo randomforest with caret

### 1.12.1 Preparación de los datos

```
betas_train <- assay(train, "counts") %>% t()
fenotipos_train <- train$label %>% factor(ordered=TRUE)

betas_test <- assay(test, "counts") %>% t()
fenotipos_test <- test$label %>% factor(ordered=TRUE)

be <- as.data.frame(betas_train)
be$label <- fenotipos_train
```

### 1.12.2 Entrenamiento del modelo

```
ctrl <- trainControl(method ="repeatedcv",
                     number=3, repeats=1,
                     selectionFunction="best",
                     savePredictions=TRUE,
                     classProbs=TRUE,
                     verboseIter=TRUE,
                     allowParallel=TRUE)

grid_rf <- expand.grid(mtry = c(100))


m_rf <- train(label ~ ., data=be, method="rf",
              trControl=ctrl,
              tuneGrid=grid_rf,
              metric="Accuracy")
```

```
## Aggregating results
## Fitting final model on full training set
```

```
m_rf
```

```
## Random Forest
##
##  7292 samples
## 10000 predictors
##    34 classes: 'ACC', 'BLCA', 'BRCA', 'CESC', 'CHOL', 'COAD', 'Control', 'DLBC', 'ESCA', 'GBM', 'HNS
##
## No pre-processing
## Resampling: Cross-Validated (3 fold, repeated 1 times)
## Summary of sample sizes: 4861, 4867, 4856
## Resampling results:
##
##   Accuracy   Kappa
##   0.9137444  0.9096866
##
## Tuning parameter 'mtry' was held constant at a value of 100
```

### 1.12.3 Evaluación del modelo

```
resultado2 <- predict(m_rf, newdata=betas_test, type="raw")

c5 <- confusionMatrix(fenotipos_test, resultado2)
c5
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction ACC BLCA BRCA CESC CHOL COAD Control DLBC ESCA GBM HNSC KICH KIRC
```

```
##     ACC       18    0    0    0    0    0        0    0    0    0    0    0    0
##     BLCA       0   96    0    0    0    0        2    1    0    0    1    0    0
##     BRCA       0    1  192    0    0    0        1    0    0    0    1    0    0
##     CESC       0    0    1   72    0    0        1    0    0    0    0    0    0
##     CHOL       0    0    0    0    4    0        1    0    0    0    0    0    0
##     COAD       0    0    0    0    0   72        0    0    0    0    0    0    0
##     Control    0    0    2    0    0    0      175    0    0    0    0    0    0
##     DLBC       0    0    0    0    0    0        0   11    0    0    0    0    0
##     ESCA       0    0    1    0    0    0        0    0    1    0   18    0    0
##     GBM        0    0    0    0    0    0        0    0    0   33    0    0    0
##     HNSC       0    1    0    0    0    0        1    0    0    0  125    0    0
##     KICH       0    0    0    0    0    0        0    0    0    0    0   15    0
##     KIRC       0    0    0    0    0    0        0    0    0    0    0    1   74
##     KIRP       0    1    0    0    0    0        1    0    0    0    0    2    1
##     LAML       0    0    0    0    0    0        0    0    0    0    0    0    0
##     LGG        0    0    0    0    0    0        0    0    0    8    0    0    0
##     LIHC       0    0    0    0    1    0        1    0    0    0    0    0    0
##     LUAD       0    0    0    0    0    0        1    0    0    0    0    0    0
##     LUSC       0    1    1    0    0    0        1    0    0    0    5    0    0
##     MESO       0    0    0    0    0    0        0    0    0    0    0    0    0
##     OV         0    0    0    0    0    0        0    0    0    0    0    0    0
##     PAAD       0    1    0    0    0    0        3    0    0    0    0    0    0
##     PCPG       0    0    0    0    0    0        2    0    0    0    0    0    0
##     PRAD       0    0    0    0    0    0        3    0    0    0    0    0    0
##     READ       0    0    0    0    0   16        0    0    0    0    0    0    0
##     SARC       0    0    0    0    0    0        0    0    0    0    0    0    0
##     SKCM       0    0    0    0    0    0        1    0    0    0    0    0    0
##     STAD       0    0    0    0    0    2        1    1    0    0    2    0    0
##     TGCT       0    0    0    0    0    0        1    0    0    0    0    0    0
##     THCA       0    0    0    0    0    0        1    0    0    0    0    0    0
##     THYM       0    0    0    0    0    0        0    0    0    0    2    0    0
##     UCEC       0    0    0    0    0    0        0    0    0    0    0    0    0
##     UCS        0    0    0    0    0    0        0    0    0    0    0    0    0
##     UVM        0    0    0    0    0    0        0    0    0    0    0    0    0
##              Reference
## Prediction KIRP LAML LGG LIHC LUAD LUSC MESO  OV PAAD PCPG PRAD READ SARC SKCM
##     ACC       0    0    0    0    0    1    0    0    0    0    0    0    1    0
##     BLCA      0    0    0    0    0    1    0    0    0    0    0    0    1    1
##     BRCA      0    0    0    0    0    0    0    0    0    0    0    0    2    0
##     CESC      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##     CHOL      0    0    0    2    1    0    0    0    0    0    0    0    0    0
##     COAD      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##     Control   0    0    0    0    1    0    0    0    0    0    4    0    0    0
##     DLBC      0    1    0    0    0    0    0    0    0    0    0    0    0    0
##     ESCA      0    0    0    0    0    4    0    0    0    0    0    0    0    0
##     GBM       0    0    5    0    0    0    0    0    0    0    0    0    0    0
##     HNSC      0    0    0    0    0    5    0    0    0    0    0    0    0    0
##     KICH      1    0    0    0    0    0    0    0    0    0    0    0    0    0
##     KIRC      2    0    0    0    0    0    0    0    0    0    0    0    3    0
##     KIRP     64    0    0    0    0    0    0    0    0    0    0    0    0    0
##     LAML      0   48    0    0    0    0    0    0    0    0    0    0    0    0
##     LGG       0    0  123    0    0    0    0    0    0    1    0    0    0    0
##     LIHC      0    0    0   90    0    0    0    0    0    0    0    0    1    1
##     LUAD      0    0    0    0  112    1    0    0    0    0    0    0    0    0
```

```
##     LUSC       0    0   0    0    7   76    0   0    0    0    0    0    0    1
##     MESO       0    0   0    0    0    0   21   0    0    0    0    0    0    0
##     OV         0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     PAAD       0    0   0    0    0    0    0   0   40    1    0    0    0    0
##     PCPG       0    0   0    0    0    0    0   0    0   44    0    0    0    0
##     PRAD       0    0   0    0    0    0    0   0    0    0  121    0    0    0
##     READ       0    0   0    0    0    0    0   0    0    0    0    8    0    0
##     SARC       0    0   0    0    0    0    0   0    0    0    0    0   65    1
##     SKCM       0    0   0    0    0    0    0   0    0    0    0    0    0  117
##     STAD       0    0   0    0    0    1    0   0    1    0    0    0    0    0
##     TGCT       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     THCA       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     THYM       0    0   0    0    0    1    0   0    0    0    0    0    0    0
##     UCEC       0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     UCS        0    0   0    0    0    0    0   0    0    0    0    0    0    0
##     UVM        0    0   0    0    0    0    0   0    0    0    0    0    0    0
##               Reference
## Prediction STAD TGCT THCA THYM UCEC UCS UVM
##     ACC       0    0    0    0    0   0   0
##     BLCA      0    0    0    0    0   0   0
##     BRCA      0    0    0    0    0   0   0
##     CESC      0    0    0    0    3   0   0
##     CHOL      1    0    0    0    0   0   0
##     COAD      1    0    0    0    0   0   0
##     Control   0    0    1    0    0   0   0
##     DLBC      0    0    0    0    0   0   0
##     ESCA     22    0    0    0    0   0   0
##     GBM       0    0    0    0    0   0   0
##     HNSC      0    0    0    0    0   0   0
##     KICH      0    0    0    0    0   0   0
##     KIRC      0    0    0    0    0   0   0
##     KIRP      0    0    0    0    0   0   0
##     LAML      0    0    0    0    0   0   0
##     LGG       0    0    0    0    0   0   0
##     LIHC      0    0    0    0    0   0   0
##     LUAD      0    0    0    0    1   0   0
##     LUSC      0    0    0    0    0   0   0
##     MESO      0    0    0    0    0   0   0
##     OV        0    0    0    0    2   0   0
##     PAAD      1    0    0    0    0   0   0
##     PCPG      0    0    0    0    0   0   0
##     PRAD      0    0    0    0    0   0   0
##     READ      0    0    0    0    0   0   0
##     SARC      0    0    0    0    0   0   0
##     SKCM      0    0    0    0    0   0   0
##     STAD     90    0    0    0    0   0   0
##     TGCT      0   33    0    0    0   0   0
##     THCA      0    0  126    0    0   0   0
##     THYM      0    0    0   28    0   0   0
##     UCEC      0    0    0    0  108   0   0
##     UCS       0    0    0    0    8   6   0
##     UVM       0    0    0    0    0   0  20
##
## Overall Statistics
```

```
##
##                Accuracy : 0.9226
##                  95% CI : (0.9112, 0.9329)
##     No Information Rate : 0.0816
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9189
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL
## Sensitivity            1.000000     0.95050     0.97462     1.00000    0.800000
## Specificity            0.999166     0.99697     0.99775     0.99787    0.997925
## Pos Pred Value         0.900000     0.93204     0.97462     0.93506    0.444444
## Neg Pred Value         1.000000     0.99784     0.99775     1.00000    0.999584
## Prevalence             0.007453     0.04182     0.08157     0.02981    0.002070
## Detection Rate         0.007453     0.03975     0.07950     0.02981    0.001656
## Detection Prevalence   0.008282     0.04265     0.08157     0.03188    0.003727
## Balanced Accuracy      0.999583     0.97373     0.98618     0.99893    0.898963
##                      Class: COAD Class: Control Class: DLBC Class: ESCA
## Sensitivity              0.80000        0.88832    0.846154   1.0000000
## Specificity              0.99957        0.99639    0.999584   0.9813587
## Pos Pred Value           0.98630        0.95628    0.916667   0.0217391
## Neg Pred Value           0.99231        0.99014    0.999168   1.0000000
## Prevalence               0.03727        0.08157    0.005383   0.0004141
## Detection Rate           0.02981        0.07246    0.004555   0.0004141
## Detection Prevalence     0.03023        0.07578    0.004969   0.0190476
## Balanced Accuracy        0.89978        0.94236    0.922869   0.9906794
##                      Class: GBM Class: HNSC Class: KICH Class: KIRC Class: KIRP
## Sensitivity             0.80488     0.81169    0.833333     0.98667     0.95522
## Specificity             0.99789     0.99690    0.999583     0.99744     0.99787
## Pos Pred Value          0.86842     0.94697    0.937500     0.92500     0.92754
## Neg Pred Value          0.99663     0.98730    0.998749     0.99957     0.99872
## Prevalence              0.01698     0.06377    0.007453     0.03106     0.02774
## Detection Rate          0.01366     0.05176    0.006211     0.03064     0.02650
## Detection Prevalence    0.01573     0.05466    0.006625     0.03313     0.02857
## Balanced Accuracy       0.90139     0.90430    0.916458     0.99205     0.97655
##                      Class: LAML Class: LGG Class: LIHC Class: LUAD Class: LUSC
## Sensitivity              0.97959    0.96094     0.97826     0.92562     0.84444
## Specificity              1.00000    0.99606     0.99828     0.99869     0.99312
## Pos Pred Value           1.00000    0.93182     0.95745     0.97391     0.82609
## Neg Pred Value           0.99958    0.99781     0.99914     0.99609     0.99397
## Prevalence               0.02029    0.05300     0.03810     0.05010     0.03727
## Detection Rate           0.01988    0.05093     0.03727     0.04638     0.03147
## Detection Prevalence     0.01988    0.05466     0.03892     0.04762     0.03810
## Balanced Accuracy        0.98980    0.97850     0.98827     0.96216     0.91878
##                      Class: MESO Class: OV Class: PAAD Class: PCPG Class: PRAD
## Sensitivity             1.000000        NA     0.97561     0.95652     0.96800
## Specificity             1.000000 0.9991718     0.99747     0.99916     0.99869
## Pos Pred Value          1.000000        NA     0.86957     0.95652     0.97581
## Neg Pred Value          1.000000        NA     0.99958     0.99916     0.99825
## Prevalence              0.008696 0.0000000     0.01698     0.01905     0.05176
```

```
## Detection Rate          0.008696 0.0000000     0.01656     0.01822     0.05010
## Detection Prevalence    0.008696 0.0008282     0.01905     0.01905     0.05135
## Balanced Accuracy       1.000000        NA     0.98654     0.97784     0.98334
##                       Class: READ Class: SARC Class: SKCM Class: STAD
## Sensitivity              1.000000     0.89041     0.96694     0.78261
## Specificity              0.993353     0.99957     0.99956     0.99652
## Pos Pred Value           0.333333     0.98485     0.99153     0.91837
## Neg Pred Value           1.000000     0.99659     0.99826     0.98921
## Prevalence               0.003313     0.03023     0.05010     0.04762
## Detection Rate           0.003313     0.02692     0.04845     0.03727
## Detection Prevalence     0.009938     0.02733     0.04886     0.04058
## Balanced Accuracy        0.996676     0.94499     0.98325     0.88957
##                       Class: TGCT Class: THCA Class: THYM Class: UCEC Class: UCS
## Sensitivity               1.00000     0.99213     1.00000     0.88525   1.000000
## Specificity               0.99958     0.99956     0.99874     1.00000   0.996679
## Pos Pred Value            0.97059     0.99213     0.90323     1.00000   0.428571
## Neg Pred Value            1.00000     0.99956     1.00000     0.99393   1.000000
## Prevalence                0.01366     0.05259     0.01159     0.05052   0.002484
## Detection Rate            0.01366     0.05217     0.01159     0.04472   0.002484
## Detection Prevalence      0.01408     0.05259     0.01284     0.04472   0.005797
## Balanced Accuracy         0.99979     0.99584     0.99937     0.94262   0.998340
##                       Class: UVM
## Sensitivity             1.000000
## Specificity             1.000000
## Pos Pred Value          1.000000
## Neg Pred Value          1.000000
## Prevalence              0.008282
## Detection Rate          0.008282
## Detection Prevalence    0.008282
## Balanced Accuracy       1.000000
```

### 1.13 SessionInfo

```
sessionInfo()
```

```
## R version 4.2.0 (2022-04-22 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Spanish_Spain.utf8  LC_CTYPE=Spanish_Spain.utf8
## [3] LC_MONETARY=Spanish_Spain.utf8 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Spain.utf8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] TCGAutils_1.16.1            doParallel_1.0.17
```

```
##  [3] iterators_1.0.14           foreach_1.5.2
##  [5] googledrive_2.0.0          randomForest_4.7-1.1
##  [7] keras_2.9.0                gmodels_2.18.1.1
##  [9] C50_0.1.6                  class_7.3-20
## [11] caret_6.0-93               lattice_0.20-45
## [13] sva_3.44.0                 BiocParallel_1.30.4
## [15] genefilter_1.78.0          mgcv_1.8-41
## [17] nlme_3.1-160               dplyr_1.0.10
## [19] ggplot2_3.3.6              Rtsne_0.16
## [21] GEOquery_2.64.2            SummarizedExperiment_1.26.1
## [23] Biobase_2.56.0             GenomicRanges_1.48.0
## [25] GenomeInfoDb_1.32.4        IRanges_2.30.1
## [27] S4Vectors_0.34.0           BiocGenerics_0.42.0
## [29] MatrixGenerics_1.8.1       matrixStats_0.62.0
## [31] knitr_1.40
##
## loaded via a namespace (and not attached):
##   [1] BiocFileCache_2.4.0         plyr_1.8.7
##   [3] splines_4.2.0               listenv_0.8.0
##   [5] tfruns_1.5.1                digest_0.6.30
##   [7] htmltools_0.5.3             gdata_2.18.0.1
##   [9] fansi_1.0.3                 magrittr_2.0.3
##  [11] memoise_2.0.1               tzdb_0.3.0
##  [13] limma_3.52.4                recipes_1.0.2
##  [15] globals_0.16.1              Biostrings_2.64.1
##  [17] readr_2.1.3                 annotate_1.74.0
##  [19] gower_1.0.0                 R.utils_2.12.0
##  [21] hardhat_1.2.0               prettyunits_1.1.1
##  [23] colorspace_2.0-3            rvest_1.0.3
##  [25] rappdirs_0.3.3              blob_1.2.3
##  [27] xfun_0.34                   crayon_1.5.2
##  [29] RCurl_1.98-1.9              jsonlite_1.8.3
##  [31] libcoin_1.0-9               zeallot_0.1.0
##  [33] survival_3.4-0              glue_1.6.2
##  [35] GenomicDataCommons_1.20.3   gargle_1.2.1
##  [37] gtable_0.3.1                ipred_0.9-13
##  [39] zlibbioc_1.42.0             XVector_0.36.0
##  [41] DelayedArray_0.22.0         future.apply_1.9.1
##  [43] scales_1.2.1                mvtnorm_1.1-3
##  [45] DBI_1.1.3                   edgeR_3.38.4
##  [47] Rcpp_1.0.9                  progress_1.2.2
##  [49] xtable_1.8-4                Cubist_0.4.0
##  [51] reticulate_1.26             proxy_0.4-27
##  [53] bit_4.0.4                   Formula_1.2-4
##  [55] lava_1.6.10                 prodlim_2019.11.13
##  [57] httr_1.4.4                  ellipsis_0.3.2
##  [59] farver_2.1.1                R.methodsS3_1.8.2
##  [61] pkgconfig_2.0.3             XML_3.99-0.11
##  [63] dbplyr_2.2.1                nnet_7.3-18
##  [65] here_1.0.1                  locfit_1.5-9.6
##  [67] utf8_1.2.2                  labeling_0.4.2
##  [69] tidyselect_1.2.0            rlang_1.0.6
##  [71] reshape2_1.4.4              AnnotationDbi_1.58.0
##  [73] munsell_0.5.0               tools_4.2.0
```

```
##  [75] cachem_1.0.6                cli_3.4.1
##  [77] generics_0.1.3              RSQLite_2.2.18
##  [79] evaluate_0.17               stringr_1.4.1
##  [81] fastmap_1.1.0               yaml_2.3.6
##  [83] fs_1.5.2                    ModelMetrics_1.2.2.2
##  [85] bit64_4.0.5                 purrr_0.3.5
##  [87] KEGGREST_1.36.3             future_1.28.0
##  [89] whisker_0.4                 R.oo_1.25.0
##  [91] xml2_1.3.3                  biomaRt_2.52.0
##  [93] compiler_4.2.0              rstudioapi_0.14
##  [95] filelock_1.0.2              curl_4.3.3
##  [97] png_0.1-7                   e1071_1.7-11
##  [99] tibble_3.1.8                stringi_1.7.8
## [101] highr_0.9                   GenomicFeatures_1.48.4
## [103] Matrix_1.5-1                tensorflow_2.9.0
## [105] vctrs_0.5.0                 pillar_1.8.1
## [107] lifecycle_1.0.3             data.table_1.14.2
## [109] bitops_1.0-7                rtracklayer_1.56.1
## [111] BiocIO_1.6.0                R6_2.5.1
## [113] parallelly_1.32.1           codetools_0.2-18
## [115] MASS_7.3-58.1               gtools_3.9.3
## [117] assertthat_0.2.1            rprojroot_2.0.3
## [119] rjson_0.2.21                withr_2.5.0
## [121] GenomicAlignments_1.32.1    Rsamtools_2.12.0
## [123] GenomeInfoDbData_1.2.8      MultiAssayExperiment_1.22.0
## [125] hms_1.1.2                   grid_4.2.0
## [127] rpart_4.1.19                timeDate_4021.106
## [129] tidyr_1.2.1                 rmarkdown_2.17
## [131] inum_1.0-4                  partykit_1.2-16
## [133] pROC_1.18.0                 lubridate_1.8.0
## [135] base64enc_0.1-3             restfulr_0.0.15
```

## Bibliografía


Kuhn, Max. 2017. *A Short Introduction to the caret Package*. https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf.

Lantz, Brett. 2015. *Machine learning with R*. Packt Publishing Ltd. http://www.packtpub.com/books/content/machine-learning-r.

Maros, Máté E, David Capper, David TW Jones, Volker Hovestadt, Andreas von Deimling, Stefan M Pfister, Axel Benner, Manuela Zucknick, y Martin Sill. 2020. «Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data». *Nature protocols* 15 (2): 479-512.

Price, E Magda, Allison M Cotton, Lucia L Lam, Pau Farré, Eldon Emberly, Carolyn J Brown, Wendy P Robinson, y Michael S Kobor. 2013. «Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array». *Epigenetics & chromatin* 6 (1): 1-15.