

Clasificador diseasecodes TCGA

Alberto Joven Álvarez

10 de noviembre, 2022

Índice

1	Clasificador machine learning de muestras tumorales.	2
1.1	Descripción datos de entrada	2
1.2	Carga inicial de los datos	2
1.3	Sondas problemáticas de acuerdo a (Price et al. 2013)	2
1.4	Gráfico previo Rtsne	4
1.5	Estimación efectos covariates (efecto batch)	6
1.6	Desglose entre muestras de entrenamiento y test	7
1.7	Algoritmo k-Nearest	8
1.8	Algoritmo árbol de decisión	13
1.9	Algoritmo red neuronal	17
1.10	Algoritmo randomforest	22
1.11	Best model randomforest con caret	26
	Bibliografía	31

```
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=4)
```

1 Clasificador machine learning de muestras tumorales.

1.1 Descripción datos de entrada

En este clasificador, como datos de entrada, se utilizan las sondas seleccionadas con los siguientes criterios: Arrays: Illumina methylation 450k.

Criterio de metilación diferencial: p-value ajustado BH para la diferencia de medias entre los grupos contrastados < 0.01 Se ha contrastado cada tumor contra el conjunto de muestras de control con la función TCGAanalyze que emplea el test de Wilcoxon, por ejemplo:

```
sondas_BRCA <- TCGAanalyze_DMC(  
  data = data,  
  groupCol = "label",  
  group1 = "Control",  
  group2 = "BRCA",  
  save.directory="G:/TFM UOC/Expresion diferencial",  
  plot.filename = "methylation_volcano_BRCA.pdf"  
)
```

Se han seleccionado las sondas que cumplían el criterio de False Discovery Rate ajustado con el criterio BH menor de 0.01, no se ha considerado un valor mínimo para la diferencia en las medias de los valores betas entre ambos grupos.

El subset de sondas se obtiene por la intersección de los subconjuntos de sondas diferenciales obtenidos con los 33 contrastes tumor vs Control.

1.2 Carga inicial de los datos

Carga del objeto **SummarizedExperiment** que contiene los valores betas, valores de fenotipos y rangos de las sondas seleccionadas:

```
# load("G:/TFM UOC/datos/data_1524sondas.Rda")  
url <- "https://drive.google.com/file/d/1mf06hvdWijT3z_hHFYEnEtAHSZ11l8po/view?usp=sharing"  
drive_download(url, overwrite = TRUE)  
  
load("data_1524sondas.Rda")  
data_1524sondas
```

```
## class: RangedSummarizedExperiment  
## dim: 1524 9707  
## metadata(0):  
## assays(1): counts  
## rownames(1524): cg00021933 cg00027990 ... ch.3.638689R ch.5.432310R  
## rowData names(10): addressA addressB ... probeEnd probeTarget  
## colnames(9707): TCGA-2F-A9K0-01A-11D-A38H-05  
##   TCGA-2F-A9KP-01A-11D-A38H-05 ... TCGA-ZA-A8F6-01A-23D-A365-05  
##   TCGA-ZQ-A9CR-01A-11D-A398-05  
## colData names(7): bar_code sujeto ... assay label
```

1.3 Sondas problemáticas de acuerdo a (Price et al. 2013)

Carga del fichero GSE42409_family.soft:

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

El porcentaje de sondas que apuntan a SNPs. Las identifico (no estoy seguro de que sea la forma correcta) con la variable: Target CpG SNP, presumo que las sondas que apuntan a SNP's son aquellas que tienen dicha variable no vacía.

```
sum(GSE42409$`Target CpG SNP` != "") * 100 / dim(GSE42409)[1]
```

```
## featureNames
##      3.992845
```

El porcentaje de sondas con más de un sitio de unión en el genoma. Las identifico si la variable AlleleA_Hits no es igual a 1 (tampoco estoy seguro de que sea la forma correcta).

```
sum(GSE42409$AlleleA_Hits != 1) * 100 / dim(GSE42409)[1]
```

```
## featureNames
##      0.1099912
```

Sondas de nuestra selección que apuntan a SNPs:

```
sondas_1524 <- GSE42409[GSE42409$ID %in% rownames(data_1524sondas), ]
sum(sondas_1524$`Target CpG SNP` != "")
```

```
## [1] 30
```

Sondas de nuestra selección con más de un sitio de unión:

```
sum(sondas_1524$AlleleA_Hits != 1)
```

```
## [1] 2
```

Suprimo estos dos tipos de sondas problemáticas:

```
s1 <- sondas_1524[sondas_1524$`Target CpG SNP` != "" | sondas_1524$AlleleA_Hits != 1,] %>% row.names()
data_sondas <- data_1524sondas[!(row.names(data_1524sondas) %in% s1)]
```

```
data_sondas
```

```
## class: RangedSummarizedExperiment
## dim: 1492 9707
## metadata(0):
## assays(1): counts
## rownames(1492): cg00021933 cg00027990 ... ch.3.638689R ch.5.432310R
## rowData names(10): addressA addressB ... probeEnd probeTarget
## colnames(9707): TCGA-2F-A9K0-01A-11D-A38H-05
##      TCGA-2F-A9KP-01A-11D-A38H-05 ... TCGA-ZA-A8F6-01A-23D-A365-05
##      TCGA-ZQ-A9CR-01A-11D-A398-05
## colData names(7): bar_code sujeto ... assay label
```

```

betas <- assay(data_sondas, "counts")
rangos_sondas <- rowRanges(data_sondas)
fenotipos <- colData(data_sondas)

etiqueta <- as.factor(fenotipos$label)

```

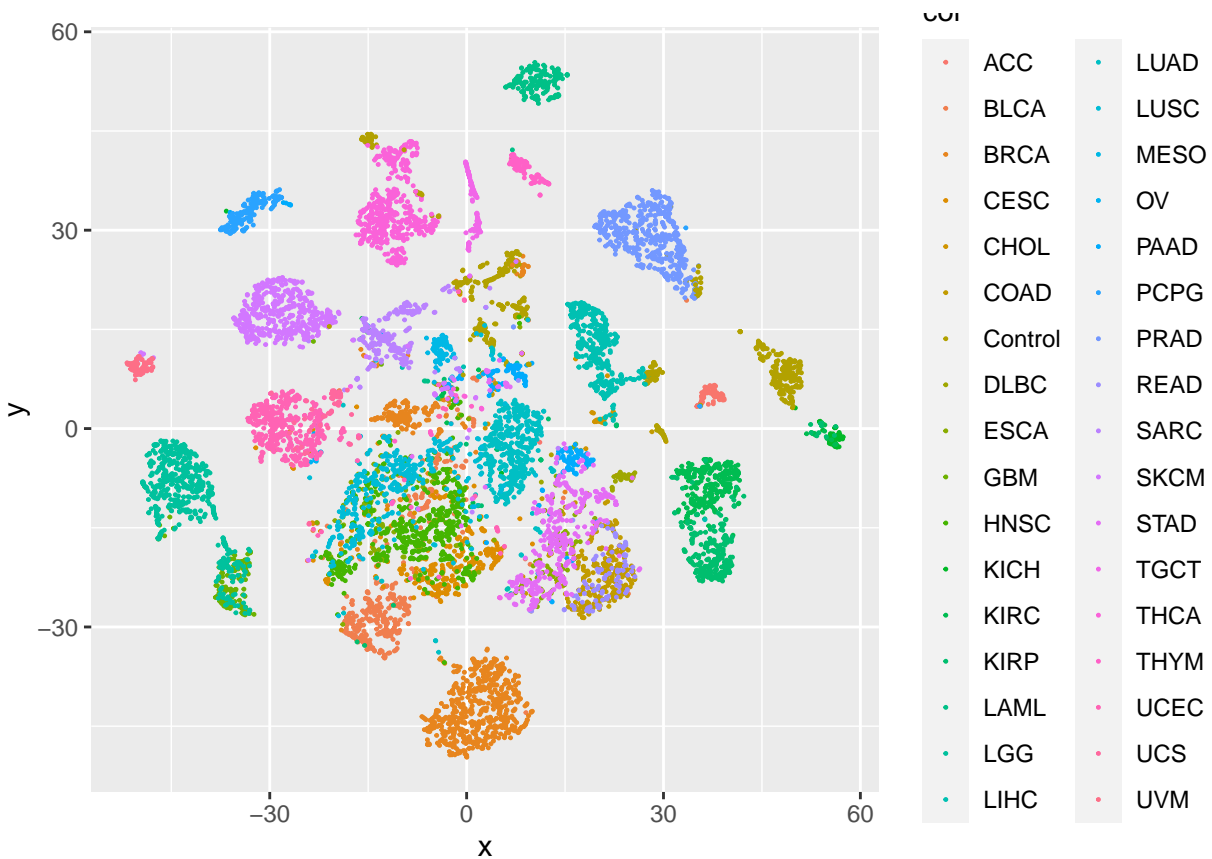
1.4 Gráfico previo Rtsne

```

sed.seed=123
tsne <- Rtsne(t(betas), 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)

```



Distribución de las sondas seleccionadas según la clase de CpG a la que apuntan:

```

t1 <- table(GSE42409$HIL_CpG_class)
t2 <- table(sondas_1524$HIL_CpG_class)
(t <- rbind(t1, t2)) %>% kable()

```

	HC	IC	ICshore	LC
t1	139826	100164	30467	157759
t2	366	370	112	538

```
t1_p <- table(GSE42409$HIL_CpG_class) %>% prop.table() %>% round(2)
t2_p <- table(sondas_1524$HIL_CpG_class) %>% prop.table() %>% round(2)
kable(rbind(t1_p, t2_p))
```

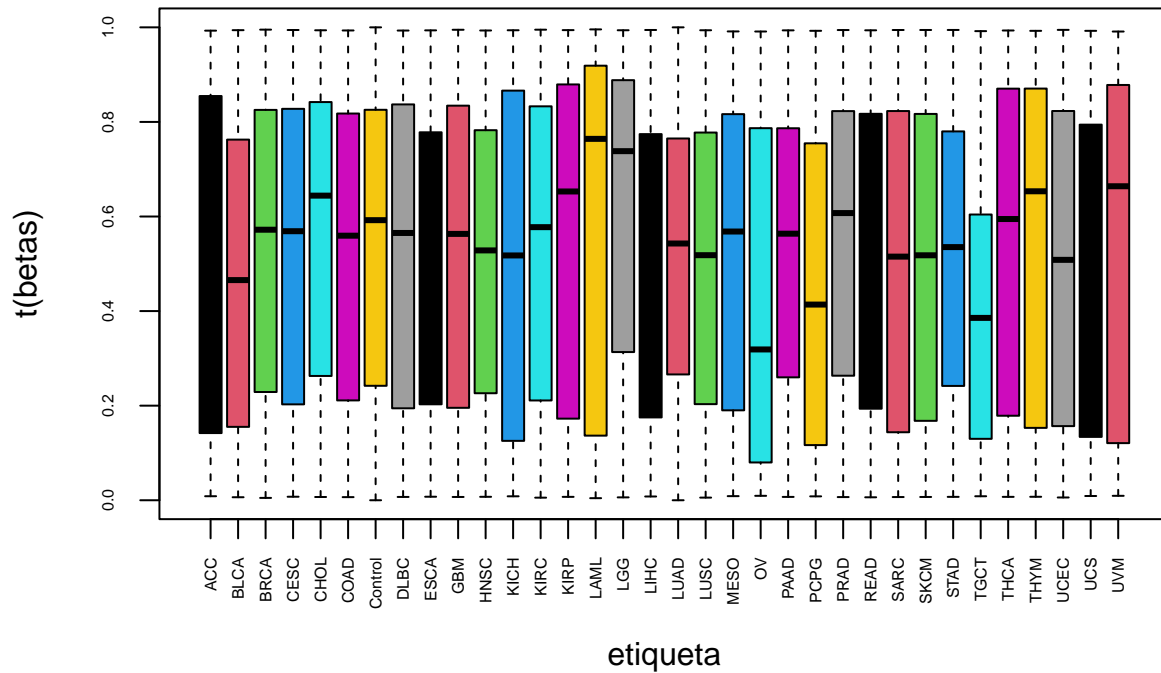
	HC	IC	ICshore	LC
t1_p	0.33	0.23	0.07	0.37
t2_p	0.26	0.27	0.08	0.39

```
chisq.test(t)
```

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 26.238, df = 3, p-value = 8.502e-06
```

El box-plot de las muestras, agrupado por tumores:

```
boxplot(t(betas) ~ etiqueta, cex.axis=0.5, las=3, col=palette())
```



1.5 Estimación efectos covariantes (efecto batch)

1.5.1 Librería sva para estimar las variables batch

Esta estimación se hace con la librería **sva**. De acuerdo con la ayuda de la función **sva**: *Esta función es la implementación del método de mínimos cuadrados ponderados iterativamente para estimar variables sustitutas.*

```
mod <- model.matrix( ~ label, data=fenotipos)
mod0 <- model.matrix( ~ 1, data=fenotipos)
```

```
n.sv = num.sv(betas, mod, method="leek")
n.sv
```

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

```
sva1 <- sva(betas, mod, mod0, n.sv=2)
```

```
## Number of significant surrogate variables is: 2
## Iteration (out of 5):1 2 3 4 5
```

```
# sale sva1$sv una matriz de 9707 x 2
# la primera columna un 1 en la primera posición y luego todos ceros
# la segunda columna un 1 en la segunda posición y luego todo ceros
sva1[[1]] %>% head()
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
## [3,]    0    0
## [4,]    0    0
## [5,]    0    0
## [6,]    0    0
```

No se detecta ningún efecto **batch** con este sistema. Otro sistema alternativo como identificar a priori las **covariates** que influyen en la variable respuesta, creo que no es aplicable en este caso, al no disponer de datos masivos de todas las muestras de los valores de pureza tumoral, centro donde se realizaron los análisis o fecha de los mismos.

1.5.2 Librería sva utilizando com variable batch el plate_id del barcode

```
nombres <- colnames(betas)
plate_id <- TCGAbiospec(nombres)
plate_id <- plate_id$plate

fenotipos$plate_id <- plate_id

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

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

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

1.6 Desglose entre muestras de entrenamiento y test

En esta primera aproximación separo el 75 % de las observaciones para entrenamiento y el 25% restante para test.

```
set.seed(123)
in_train <- createDataPartition(etiqueta, p=0.75, list=FALSE)

train <- data_sondas[ , as.vector(in_train)]
test <- data_sondas[ , as.vector(-in_train)]

df <- data.frame(Train= table(colData(train)$label),
                Test = table(colData(test)$label),
                Total = table(colData(data_sondas)$label))

df[, c(1,2,4,6) ] %>% kable()
```

Train.Var1	Train.Freq	Test.Freq	Total.Freq
ACC	60	20	80
BLCA	310	103	413

Train.Var1	Train.Freq	Test.Freq	Total.Freq
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
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

```

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

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

```

1.7 Algoritmo k-Nearest

Este algoritmo asigna la clase **Tipo tumor** de las observaciones test a la clase mayoritaria de las k observaciones más cercanas.

Estimación del modelo:

Inicialmente probamos con un modelo con el valor del parámetro k que indica el número de observaciones vecinas a considerar en la clasificación, igual a 5.

Se utiliza la función `knn` del paquete `class`.


```
modelo_knn <- class::knn(betas_train, betas_test, cl=fenotipos_train, k=10)
```

Evaluación del modelo

Se utiliza la función confusionMatrix del paquete caret para evaluar el modelo con los datos del grupo de observaciones test.

```
c1 <- confusionMatrix(modelo_knn, fenotipos_test, positive="Control" )
c1
```

```
## Confusion Matrix and Statistics
```

```
##
```

```
##           Reference
```

```
## Prediction ACC BLCA BRCA CESC CHOL COAD Control DLBC ESCA GBM HNSC KICH KIRC
```

```
## ACC        6    0    0    0    0    0    0    0    0    0    0    0    0    0
```

```
## BLCA       0   78    0    1    0    0    1    0    1    0    1    0    0    1
```

```
## BRCA       0    0   171    0    0    0    0    0    0    0    0    0    0    0
```

```
## CESC       0    0    0   38    0    0    0    0    0    0    0    4    0    0
```

```
## CHOL       0    0    0    0    5    0    0    0    0    0    0    0    0    0
```

```
## COAD       0    0    0    0    0   64    0    0    1    0    0    0    0    0
```

```
## Control    4    2   12    2    0    0   177    0    1    1    1   15   10
```

```
## DLBC       0    0    0    0    0    0    0    4    0    0    0    0    0    0
```

```
## ESCA       0    1    0    0    0    0    0    0    6    0    0    0    0    0
```

```
## GBM        0    0    0    0    0    0    0    0    0    27    0    0    0    0
```

```
## HNSC       1    9    4   25    0    0    0    0   13    1  113    1    2
```

```
## KICH       0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

```
## KIRC       0    0    0    0    0    0    0    0    0    0    2    0    0   59
```

```
## KIRP       0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

```
## LAML       0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

```
## LGG        0    0    0    0    0    0    0    0    0    0    5    0    0    0
```

```
## LIHC       1    0    0    0    1    0    0    0    0    0    0    0    0    0
```

```
## LUAD       5    4    6    2    1    1    0    4    1    1    0    0    0    6
```

```
## LUSC       0    4    2    4    0    0    0    0    9    1   12    0    0    0
```

```
## MESO       0    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    0
```

```
## PAAD       0    1    0    1    1    0    1    1    0    0    0    0    0    1
```

```
## PCPG       0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

```
## PRAD       0    0    0    0    0    0    3    0    0    0    0    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    0    1
```

```
## SKCM       0    1    0    0    0    0    0    1    0    0    0    0    0    0
```

```
## STAD       0    2    1    2    1    8    1    2   14    0    0    0    0    0
```

```
## TGCT       3    0    0    0    0    0    0    0    0    0    1    0    0    0
```

```
## THCA       0    1    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    2    0    0    0    0    0    0    0    0    0    0
```

```
## UCS        0    0    1    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 KIRP LAML LGG LIHC LUAD LUSC MESO OV PAAD PCPG PRAD READ SARC SKCM
```

```
## ACC        0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

```
## BLCA       1    0    0    0    2    0    0    0    0    0    0    0    1    0
```

```
## BRCA       0    0    0    0    0    1    0    0    0    0    0    0    1    0
```

##	CESC	0	0	0	0	0	0	0	0	0	0	0	0	0	
##	CHOL	0	0	0	1	0	0	0	0	0	0	0	0	0	
##	COAD	0	0	0	0	0	0	0	0	0	0	23	0	0	
##	Control	4	1	1	6	3	1	1	0	2	13	4	0	14	3
##	DLBC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	ESCA	0	0	0	0	0	1	0	0	0	0	0	0	0	0
##	GBM	0	0	1	0	0	0	0	0	0	0	0	0	0	0
##	HNSC	0	1	0	1	0	12	0	0	1	0	0	0	3	8
##	KICH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	KIRC	10	0	0	0	0	0	0	0	0	0	0	0	2	0
##	KIRP	50	0	0	0	0	0	0	0	0	0	0	0	0	0
##	LAML	0	30	0	0	0	0	0	0	0	0	0	0	0	0
##	LGG	0	0	130	0	0	0	0	0	0	0	0	0	0	0
##	LIHC	0	0	0	78	0	0	0	0	0	1	0	0	0	0
##	LUAD	4	7	0	5	105	7	0	0	0	1	0	0	2	5
##	LUSC	0	0	0	0	3	69	0	0	0	0	0	0	0	0
##	MESO	0	0	0	0	0	0	20	0	0	0	0	0	2	0
##	OV	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	PAAD	0	1	0	1	0	1	0	0	41	1	2	0	1	0
##	PCPG	0	0	0	0	0	0	0	0	0	27	0	0	0	0
##	PRAD	0	0	0	0	0	0	0	0	0	0	118	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	38	0
##	SKCM	0	0	0	0	0	0	0	0	0	0	0	0	1	100
##	STAD	0	0	0	1	1	0	0	0	2	0	0	1	0	0
##	TGCT	0	2	0	1	0	0	0	0	0	3	0	0	1	1
##	THCA	0	1	0	0	0	0	0	0	0	0	0	0	0	1
##	THYM	0	5	0	0	0	0	0	0	0	0	0	0	0	0
##	UCEC	0	0	0	0	0	0	0	2	0	0	0	0	0	0
##	UCS	0	0	0	0	1	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	1	0	0	0	0	0	0
##	BRCA	0	0	0	0	0	0	0
##	CESC	0	0	0	0	0	0	0
##	CHOL	0	0	0	0	0	0	0
##	COAD	4	0	0	0	0	0	0
##	Control	1	0	22	1	0	0	4
##	DLBC	1	0	0	0	0	0	0
##	ESCA	4	0	0	0	0	0	0
##	GBM	0	0	0	0	0	0	0
##	HNSC	2	0	0	2	2	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	4	0	4	2	3	1	0
##	LUSC	0	0	0	1	1	0	0
##	MESO	0	0	0	0	0	0	0
##	OV	0	0	0	0	0	0	0

```

##      PAAD      3      0      0      0      1      0      0
##      PCPG      0      0      0      0      0      0      0
##      PRAD      0      0      0      1      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      77      0      0      0      0      0      0
##      TGCT      1     34      0      0      0      0      0
##      THCA      0      0    101      0      0      0      0
##      THYM      0      0      0     24      0      0      0
##      UCEC      0      0      0      0    101      8      0
##      UCS       0      0      0      0      0      5      0
##      UVM       0      0      0      0      0      0     16
##
## Overall Statistics
##
##           Accuracy : 0.7917
##           95% CI : (0.775, 0.8078)
##           No Information Rate : 0.0816
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.7813
##
## Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##           Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL
## Sensitivity      0.300000      0.75728      0.86802      0.49351      0.555556
## Specificity      1.000000      0.99567      0.99910      0.99829      0.999584
## Pos Pred Value   1.000000      0.88636      0.98844      0.90476      0.833333
## Neg Pred Value   0.994188      0.98926      0.98840      0.98357      0.998340
## Prevalence       0.008282      0.04265      0.08157      0.03188      0.003727
## Detection Rate   0.002484      0.03230      0.07081      0.01573      0.002070
## Detection Prevalence 0.002484      0.03644      0.07164      0.01739      0.002484
## Balanced Accuracy 0.650000      0.87648      0.93356      0.74590      0.777570
##
##           Class: COAD Class: Control Class: DLBC Class: ESCA
## Sensitivity      0.87671      0.96721      0.333333      0.130435
## Specificity      0.98804      0.94220      0.999584      0.997467
## Pos Pred Value   0.69565      0.57843      0.800000      0.500000
## Neg Pred Value   0.99613      0.99716      0.996680      0.983354
## Prevalence       0.03023      0.07578      0.004969      0.019048
## Detection Rate   0.02650      0.07329      0.001656      0.002484
## Detection Prevalence 0.03810      0.12671      0.002070      0.004969
## Balanced Accuracy 0.93238      0.95471      0.666459      0.563951
##
##           Class: GBM Class: HNSC Class: KICH Class: KIRC Class: KIRP
## Sensitivity      0.71053      0.85606      0.000000      0.73750      0.72464
## Specificity      0.99958      0.96145      1.000000      0.99400      1.00000
## Pos Pred Value   0.96429      0.56219      NaN          0.80822      1.00000
## Neg Pred Value   0.99539      0.99142      0.993375      0.99103      0.99197
## Prevalence       0.01573      0.05466      0.006625      0.03313      0.02857
## Detection Rate   0.01118      0.04679      0.000000      0.02443      0.02070
## Detection Prevalence 0.01159      0.08323      0.000000      0.03023      0.02070
## Balanced Accuracy 0.85505      0.90876      0.500000      0.86575      0.86232

```

##	Class: LAML	Class: LGG	Class: LIHC	Class: LUAD	Class: LUSC
## Sensitivity	0.62500	0.98485	0.82979	0.91304	0.75000
## Specificity	1.00000	0.99781	0.99871	0.96696	0.98407
## Pos Pred Value	1.00000	0.96296	0.96296	0.58011	0.65094
## Neg Pred Value	0.99245	0.99912	0.99314	0.99552	0.99004
## Prevalence	0.01988	0.05466	0.03892	0.04762	0.03810
## Detection Rate	0.01242	0.05383	0.03230	0.04348	0.02857
## Detection Prevalence	0.01242	0.05590	0.03354	0.07495	0.04389
## Balanced Accuracy	0.81250	0.99133	0.91425	0.94000	0.86704
##	Class: MESO	Class: OV	Class: PAAD	Class: PCPG	Class: PRAD
## Sensitivity	0.952381	0.000000	0.89130	0.58696	0.95161
## Specificity	0.999165	1.000000	0.99282	1.00000	0.99825
## Pos Pred Value	0.909091	NaN	0.70690	1.00000	0.96721
## Neg Pred Value	0.999582	0.9991718	0.99788	0.99204	0.99738
## Prevalence	0.008696	0.0008282	0.01905	0.01905	0.05135
## Detection Rate	0.008282	0.0000000	0.01698	0.01118	0.04886
## Detection Prevalence	0.009110	0.0000000	0.02402	0.01118	0.05052
## Balanced Accuracy	0.975773	0.5000000	0.94206	0.79348	0.97493
##	Class: READ	Class: SARC	Class: SKCM	Class: STAD	
## Sensitivity	0.000000	0.57576	0.84746	0.78571	
## Specificity	1.000000	0.99957	0.99869	0.98446	
## Pos Pred Value	NaN	0.97436	0.97087	0.68142	
## Neg Pred Value	0.990062	0.98822	0.99221	0.99088	
## Prevalence	0.009938	0.02733	0.04886	0.04058	
## Detection Rate	0.000000	0.01573	0.04141	0.03188	
## Detection Prevalence	0.000000	0.01615	0.04265	0.04679	
## Balanced Accuracy	0.500000	0.78767	0.92308	0.88509	
##	Class: TGCT	Class: THCA	Class: THYM	Class: UCEC	Class: UCS
## Sensitivity	1.00000	0.79528	0.774194	0.93519	0.357143
## Specificity	0.99454	0.99869	0.997903	0.99480	0.999167
## Pos Pred Value	0.72340	0.97115	0.827586	0.89381	0.714286
## Neg Pred Value	1.00000	0.98875	0.997066	0.99696	0.996262
## Prevalence	0.01408	0.05259	0.012836	0.04472	0.005797
## Detection Rate	0.01408	0.04182	0.009938	0.04182	0.002070
## Detection Prevalence	0.01946	0.04306	0.012008	0.04679	0.002899
## Balanced Accuracy	0.99727	0.89698	0.886048	0.96499	0.678155
##	Class: UVM				
## Sensitivity	0.800000				
## Specificity	1.000000				
## Pos Pred Value	1.000000				
## Neg Pred Value	0.998333				
## Prevalence	0.008282				
## Detection Rate	0.006625				
## Detection Prevalence	0.006625				
## Balanced Accuracy	0.900000				

El modelo resultante da un valor de *accuracy* de 0.79 y del parámetro Kappa de 0.78 que de acuerdo con el texto Lantz (2015), es considerado como excelente. El valor de la *sensitivity* del modelo es 0.3, este es el tanto por uno de casos malignos que han sido detectados por el algoritmo.

1.8 Algoritmo árbol de decisión

```
modelo_arbol <- C5.0(betas_train, fenotipos_train, trials=10)
modelo_arbol
```

```
##
## Call:
## C5.0.default(x = betas_train, y = fenotipos_train, trials = 10)
##
## Classification Tree
## Number of samples: 7292
## Number of predictors: 1492
##
## Number of boosting iterations: 10
## Average tree size: 397.8
##
## Non-standard options: attempt to group attributes
```

```
prediccion <- predict(modelo_arbol, betas_test)
c2 <- confusionMatrix(fenotipos_test, prediccion )
c2
```

```
## Confusion Matrix and Statistics
```

```
##
##              Reference
## Prediction ACC BLCA BRCA CESC CHOL COAD Control DLBC ESCA GBM HNSC KICH KIRC
## ACC          20   0   0   0   0   0   0   0   0   0   0   0   0   0
## BLCA          0  86   2   1   0   0   0   0   1   0   0   4   0   0
## BRCA          0   0 188   0   0   0   0   3   0   0   0   1   0   0
## CESC          0   2   1  61   0   0   0   1   0   0   0   7   0   0
## CHOL          0   0   1   0   6   0   0   1   0   0   0   0   0   0
## COAD          0   1   0   0   0   68   0   0   0   0   0   0   0   0
## Control       0   0   3   0   0   0   167   0   1   1   2   0   0   0
## DLBC          0   0   0   0   0   0   0   12   0   0   0   0   0   0
## ESCA          0   1   1   2   0   0   0   1   29   0   4   0   0   0
## GBM           0   0   0   0   0   0   0   0   0   34   0   0   0   0
## HNSC          0   3   1   3   0   0   0   1   0   2   0 117   0   0
## KICH          0   1   0   0   0   0   0   0   0   0   0   0 15   0
## KIRC          0   0   0   0   0   0   0   0   0   0   0   0   0 74
## KIRP          0   1   1   0   1   0   0   1   0   0   0   0   0   3
## LAML          0   0   0   0   0   0   0   1   0   0   0   0   0   0
## LGG           0   0   0   0   0   0   0   0   0   0   1   0   0   0
## LIHC          0   0   0   0   0   0   0   1   0   0   0   0   0   0
## LUAD          0   0   0   0   0   0   0   3   0   0   0   1   0   0
## LUSC          0   7   4   0   0   0   0   1   0   0   0   5   0   0
## MESO          0   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   0
## PAAD          0   1   0   0   0   0   0   3   0   0   0   2   0   0
## PCPG          0   0   0   0   0   0   0   1   0   0   0   0   0   0
## PRAD          0   0   2   0   0   0   0   1   0   0   0   0   0   0
## READ          0   0   0   0   0   10   0   0   0   0   0   0   0   0
## SARC          0   0   2   0   0   0   0   0   0   0   0   0   0   0
```

##	SKCM	0	1	0	0	0	0	1	0	0	0	0	0	0	
##	STAD	0	1	2	0	0	2	2	1	7	0	0	0	0	
##	TGCT	0	0	0	0	0	0	0	0	0	0	0	0	0	
##	THCA	0	0	1	0	0	0	3	0	0	0	0	0	0	
##	THYM	0	0	0	0	0	0	0	0	0	0	0	0	0	
##	UCEC	0	0	0	1	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	0	0	0	0	0	0	0	0	0
##	BLCA	0	0	0	0	0	3	1	0	0	0	1	0	0	1
##	BRCA	0	0	0	0	1	1	0	0	0	0	0	0	2	0
##	CESC	0	0	0	0	0	1	0	0	0	0	2	0	0	0
##	CHOL	0	0	0	1	0	0	0	0	0	0	0	0	0	0
##	COAD	0	0	0	1	0	0	0	0	0	0	0	3	0	0
##	Control	0	0	0	0	0	2	0	0	0	0	4	0	1	0
##	DLBC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	ESCA	0	0	0	0	1	2	0	0	1	0	0	0	0	0
##	GBM	0	0	3	0	0	0	0	0	0	0	0	0	1	0
##	HNSC	0	0	0	0	0	3	0	0	0	1	0	0	0	1
##	KICH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	KIRC	1	0	0	1	0	0	1	0	0	0	0	0	2	0
##	KIRP	62	0	0	0	0	0	0	0	0	0	0	0	0	0
##	LAML	0	47	0	0	0	0	0	0	0	0	0	0	0	0
##	LGG	1	0	130	0	0	0	0	0	0	0	0	0	0	0
##	LIHC	0	0	0	92	0	0	0	0	0	0	0	0	1	0
##	LUAD	0	0	0	0	110	0	0	0	0	0	0	0	0	0
##	LUSC	1	0	0	0	1	71	1	0	0	0	0	0	0	0
##	MESO	0	0	0	0	0	0	20	0	0	0	0	0	1	0
##	OV	0	0	0	0	0	0	0	1	0	0	0	0	0	0
##	PAAD	0	0	0	1	0	0	0	0	34	0	0	0	0	0
##	PCPG	0	0	0	0	0	0	0	0	0	45	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	14	0	0
##	SARC	0	0	0	0	0	0	0	0	0	0	0	0	60	3
##	SKCM	0	0	0	1	0	0	0	0	0	0	0	0	2	113
##	STAD	0	0	0	0	2	1	1	0	2	0	0	0	0	1
##	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	1
##	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	1	0	0	2	0	0	0							
##	BRCA	0	0	0	0	0	1	0							
##	CESC	0	0	0	0	2	0	0							
##	CHOL	0	0	0	0	0	0	0							
##	COAD	0	0	0	0	0	0	0							
##	Control	0	0	2	0	0	0	0							
##	DLBC	0	0	0	0	0	0	0							

##	ESCA	4	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	1	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	1	0	0	0	0	0	0
##	MESO	0	0	0	0	0	0	0
##	OV	0	0	0	0	1	0	0
##	PAAD	5	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	1	0	0	0	0	0	0
##	SKCM	0	0	0	0	0	0	0
##	STAD	76	0	0	0	0	0	0
##	TGCT	0	34	0	0	0	0	0
##	THCA	0	0	122	0	0	0	0
##	THYM	0	0	0	30	1	0	0
##	UCEC	2	0	0	0	104	1	0
##	UCS	0	0	0	0	0	14	0
##	UVM	0	0	0	0	0	0	20

##

Overall Statistics

##

Accuracy : 0.9097

95% CI : (0.8976, 0.9209)

No Information Rate : 0.0865

P-Value [Acc > NIR] : < 2.2e-16

##

Kappa : 0.9055

##

McNemar's Test P-Value : NA

##

Statistics by Class:

##

Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL

Sensitivity 1.000000 0.81905 0.89952 0.89706 0.857143

Specificity 1.000000 0.99264 0.99592 0.99318 0.998754

Pos Pred Value 1.000000 0.83495 0.95431 0.79221 0.666667

Neg Pred Value 1.000000 0.99178 0.99053 0.99701 0.999584

Prevalence 0.008282 0.04348 0.08654 0.02816 0.002899

Detection Rate 0.008282 0.03561 0.07785 0.02526 0.002484

Detection Prevalence 0.008282 0.04265 0.08157 0.03188 0.003727

Balanced Accuracy 1.000000 0.90584 0.94772 0.94512 0.927949

##

Class: COAD Class: Control Class: DLBC Class: ESCA

Sensitivity 0.85000 0.86979 0.857143 0.74359

Specificity 0.99786 0.99280 1.000000 0.99285

Pos Pred Value 0.93151 0.91257 1.000000 0.63043

Neg Pred Value 0.99488 0.98880 0.999168 0.99578

## Prevalence	0.03313	0.07950	0.005797	0.01615	
## Detection Rate	0.02816	0.06915	0.004969	0.01201	
## Detection Prevalence	0.03023	0.07578	0.004969	0.01905	
## Balanced Accuracy	0.92393	0.93130	0.928571	0.86822	
##	Class: GBM	Class: HNSC	Class: KICH	Class: KIRC	Class: KIRP
## Sensitivity	0.94444	0.81818	1.000000	0.96104	0.95385
## Specificity	0.99832	0.99340	0.999583	0.99743	0.99702
## Pos Pred Value	0.89474	0.88636	0.937500	0.92500	0.89855
## Neg Pred Value	0.99916	0.98861	1.000000	0.99872	0.99872
## Prevalence	0.01491	0.05921	0.006211	0.03188	0.02692
## Detection Rate	0.01408	0.04845	0.006211	0.03064	0.02567
## Detection Prevalence	0.01573	0.05466	0.006625	0.03313	0.02857
## Balanced Accuracy	0.97138	0.90579	0.999792	0.97924	0.97543
##	Class: LAML	Class: LGG	Class: LIHC	Class: LUAD	Class: LUSC
## Sensitivity	1.00000	0.97744	0.94845	0.95652	0.84524
## Specificity	0.99958	0.99912	0.99914	0.99783	0.99099
## Pos Pred Value	0.97917	0.98485	0.97872	0.95652	0.77174
## Neg Pred Value	1.00000	0.99869	0.99785	0.99783	0.99440
## Prevalence	0.01946	0.05507	0.04017	0.04762	0.03478
## Detection Rate	0.01946	0.05383	0.03810	0.04555	0.02940
## Detection Prevalence	0.01988	0.05466	0.03892	0.04762	0.03810
## Balanced Accuracy	0.99979	0.98828	0.97380	0.97717	0.91811
##	Class: MESO	Class: OV	Class: PAAD	Class: PCPG	Class: PRAD
## Sensitivity	0.833333	1.000000	0.91892	0.97826	0.94531
## Specificity	0.999582	0.9995857	0.99495	0.99958	0.99869
## Pos Pred Value	0.952381	0.500000	0.73913	0.97826	0.97581
## Neg Pred Value	0.998329	1.000000	0.99873	0.99958	0.99694
## Prevalence	0.009938	0.0004141	0.01532	0.01905	0.05300
## Detection Rate	0.008282	0.0004141	0.01408	0.01863	0.05010
## Detection Prevalence	0.008696	0.0008282	0.01905	0.01905	0.05135
## Balanced Accuracy	0.916458	0.9997929	0.95694	0.98892	0.97200
##	Class: READ	Class: SARC	Class: SKCM	Class: STAD	
## Sensitivity	0.823529	0.85714	0.94167	0.84444	
## Specificity	0.995830	0.99744	0.99782	0.99054	
## Pos Pred Value	0.583333	0.90909	0.95763	0.77551	
## Neg Pred Value	0.998745	0.99574	0.99695	0.99396	
## Prevalence	0.007039	0.02899	0.04969	0.03727	
## Detection Rate	0.005797	0.02484	0.04679	0.03147	
## Detection Prevalence	0.009938	0.02733	0.04886	0.04058	
## Balanced Accuracy	0.909680	0.92729	0.96974	0.91749	
##	Class: TGCT	Class: THCA	Class: THYM	Class: UCEC	Class: UCS
## Sensitivity	1.00000	0.98387	0.90909	0.95413	0.875000
## Specificity	1.00000	0.99782	0.99958	0.99827	1.000000
## Pos Pred Value	1.00000	0.96063	0.96774	0.96296	1.000000
## Neg Pred Value	1.00000	0.99913	0.99874	0.99783	0.999167
## Prevalence	0.01408	0.05135	0.01366	0.04513	0.006625
## Detection Rate	0.01408	0.05052	0.01242	0.04306	0.005797
## Detection Prevalence	0.01408	0.05259	0.01284	0.04472	0.005797
## Balanced Accuracy	1.00000	0.99084	0.95434	0.97620	0.937500
##	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
```

1.9 Algoritmo red neuronal

```
red <- keras_model_sequential() %>%
  layer_dense(units = 2000, activation="relu", input_shape=1492) %>%
  layer_dense(units = 35, activation = "softmax")
```

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

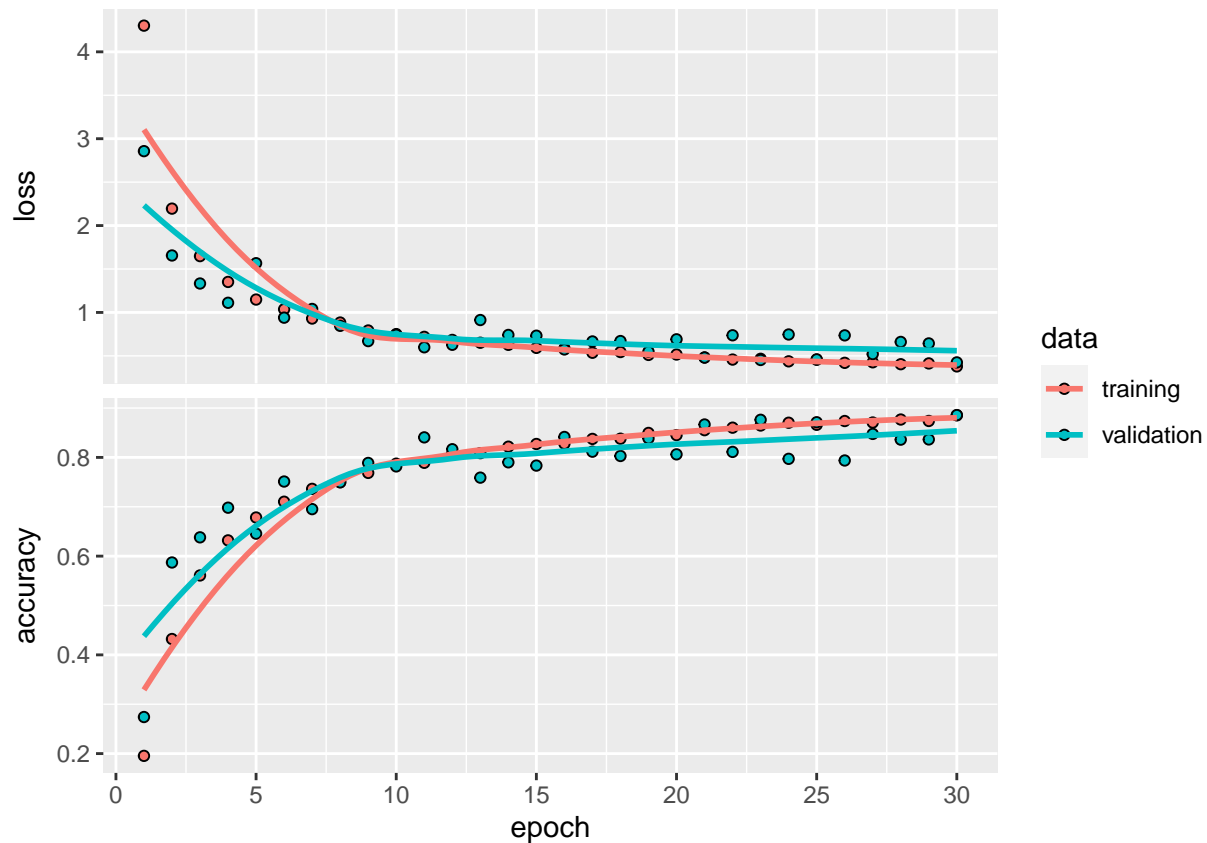
```
red %>% compile(
  optimizer = "rmsprop",
  loss = "categorical_crossentropy",
  metrics = c("accuracy")
)

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_1 (Dense)             (None, 2000)          2986000
## dense (Dense)                (None, 35)            70035
## -----
## Total params: 3,056,035
## Trainable params: 3,056,035
## Non-trainable params: 0
## -----
```

```
hist <- red %>% fit(betas_train, train_labels,
  epoch=30, batch_size=128,
  validation_data = list(betas_test, test_labels))
plot(hist)
```



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

```
##      loss  accuracy
## 0.4248793 0.8857143
```

```
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      15    0    0    0    0    0    0    0    0    0    0    0    1
##   BLCA      0   96    0    0    0    0    1    1    0    0    0    0    0
##   BRCA      0    0  192    0    0    0    2    0    0    0    0    0    0
```

##	CESC	0	3	0	64	0	0	1	0	0	0	5	0	0	
##	CHOL	0	0	0	1	6	0	0	0	0	0	0	0	0	
##	COAD	0	0	0	0	0	71	0	0	0	0	0	0	0	
##	Control	0	0	6	0	0	1	160	0	0	0	4	0	0	
##	DLBC	0	0	0	0	0	0	0	8	0	0	0	0	0	
##	ESCA	0	1	1	0	0	0	1	0	7	0	20	0	0	
##	GBM	0	0	0	0	0	0	0	0	0	32	0	0	0	
##	HNSC	0	2	0	2	0	0	1	0	0	0	121	0	0	
##	KICH	0	0	0	0	0	0	1	0	0	0	0	14	1	
##	KIRC	0	0	0	0	0	0	0	0	0	0	0	1	70	
##	KIRP	0	1	0	0	0	0	0	0	0	0	0	2	5	
##	LAML	0	0	1	0	0	0	0	0	0	0	0	0	1	
##	LGG	0	0	0	0	0	0	1	0	0	6	0	0	0	
##	LIHC	0	0	0	0	1	0	0	0	0	0	0	0	0	
##	LUAD	0	0	0	0	0	0	1	0	0	0	0	0	0	
##	LUSC	0	2	0	0	0	0	0	0	0	0	10	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	1	3	0	0	0	0	0	0	
##	PCPG	1	0	0	0	0	0	0	0	0	0	0	0	0	
##	PRAD	0	0	0	0	0	0	0	0	0	0	0	0	0	
##	READ	0	0	0	0	0	23	0	0	0	0	0	0	0	
##	SARC	0	0	0	0	0	0	0	0	0	0	0	0	0	
##	SKCM	1	0	0	0	0	0	3	0	0	0	0	0	0	
##	STAD	0	0	2	0	0	8	1	2	8	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	1	0	0	0	0	0	0	
##	THYM	0	0	2	0	0	0	0	0	0	0	0	0	0	
##	UCEC	0	0	1	2	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	1	1	0	0	0	1	0	0	0	0	0
##	BLCA	1	0	0	0	0	1	0	0	0	0	0	0	1	2
##	BRCA	0	0	0	0	0	0	0	0	0	0	0	0	2	1
##	CESC	0	0	0	0	0	1	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	2	2	1	0	0	0	0	5	0	1	1
##	DLBC	0	0	0	1	1	0	0	0	0	0	0	0	0	2
##	ESCA	0	0	0	1	0	1	0	0	0	0	0	0	0	0
##	GBM	0	0	4	0	0	0	0	0	0	0	0	0	1	1
##	HNSC	0	0	0	0	0	6	0	0	0	0	0	0	0	0
##	KICH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	KIRC	1	0	0	0	0	0	0	0	0	0	0	0	4	3
##	KIRP	60	0	0	1	0	0	0	0	0	0	0	0	0	0
##	LAML	0	44	0	0	0	0	0	0	0	0	0	0	0	2
##	LGG	0	0	124	0	0	0	0	0	0	0	0	0	0	0
##	LIHC	0	0	0	88	0	0	0	0	0	1	0	0	1	2
##	LUAD	0	0	0	0	111	3	0	0	0	0	0	0	0	0
##	LUSC	0	0	0	0	10	69	0	0	0	0	0	0	0	1
##	MESO	0	0	0	0	0	0	20	0	0	0	0	0	1	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	39	2	0	0	0	1
##	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	124	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	62	4
##	SKCM	0	0	0	0	0	0	0	0	0	0	0	0	1	113
##	STAD	0	0	0	2	2	1	0	0	1	0	0	0	0	2
##	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	1	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	1
##	UVM	0	0	0	0	0	0	0	0	0	0	0	0	0	5

##	Reference							
##	Prediction	STAD	TGCT	THCA	THYM	UCEC	UCS	UVM
##	ACC	0	1	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	0	0	0	0	0	0	0
##	COAD	2	0	0	0	0	0	0
##	Control	0	0	0	0	0	0	0
##	DLBC	0	0	0	0	0	0	0
##	ESCA	14	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	1	0
##	KIRP	0	0	0	0	0	0	0
##	LAML	0	0	0	0	0	0	0
##	LGG	0	0	0	0	0	0	1
##	LIHC	0	1	0	0	0	0	0
##	LUAD	0	0	0	0	0	0	0
##	LUSC	0	0	0	0	0	0	0
##	MESO	0	0	0	0	0	0	0
##	OV	0	0	0	0	1	0	0
##	PAAD	0	0	0	0	0	0	0
##	PCPG	0	1	0	0	0	0	0
##	PRAD	0	0	0	0	0	0	0
##	READ	1	0	0	0	0	0	0
##	SARC	0	0	0	0	0	0	0
##	SKCM	0	0	0	0	0	0	0
##	STAD	68	1	0	0	0	0	0
##	TGCT	0	34	0	0	0	0	0
##	THCA	0	0	126	0	0	0	0
##	THYM	0	0	0	27	0	0	0
##	UCEC	0	1	0	0	102	2	0
##	UCS	0	0	0	0	1	12	0
##	UVM	0	0	0	0	0	0	15

Overall Statistics

##	Accuracy :	0.8857
##	95% CI :	(0.8723, 0.8981)

```

##      No Information Rate : 0.0849
##      P-Value [Acc > NIR] : < 2.2e-16
##
##      Kappa : 0.8803
##
##      McNemar's Test P-Value : NA
##
## Statistics by Class:
##
##      Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL
## Sensitivity      0.882353      0.91429      0.93659      0.92754      0.857143
## Specificity      0.997915      0.99697      0.99774      0.99446      0.998754
## Pos Pred Value   0.750000      0.93204      0.97462      0.83117      0.666667
## Neg Pred Value   0.999165      0.99611      0.99414      0.99786      0.999584
## Prevalence       0.007039      0.04348      0.08489      0.02857      0.002899
## Detection Rate   0.006211      0.03975      0.07950      0.02650      0.002484
## Detection Prevalence 0.008282      0.04265      0.08157      0.03188      0.003727
## Balanced Accuracy 0.940134      0.95563      0.96716      0.96100      0.927949
##
##      Class: COAD Class: Control Class: DLBC Class: ESCA
## Sensitivity      0.68269      0.90395      0.727273      0.466667
## Specificity      0.99913      0.98972      0.998336      0.983750
## Pos Pred Value   0.97260      0.87432      0.666667      0.152174
## Neg Pred Value   0.98591      0.99238      0.998752      0.996623
## Prevalence       0.04306      0.07329      0.004555      0.006211
## Detection Rate   0.02940      0.06625      0.003313      0.002899
## Detection Prevalence 0.03023      0.07578      0.004969      0.019048
## Balanced Accuracy 0.84091      0.94684      0.862804      0.725208
##
##      Class: GBM Class: HNSC Class: KICH Class: KIRC Class: KIRP
## Sensitivity      0.84211      0.75625      0.823529      0.89744      0.96774
## Specificity      0.99748      0.99512      0.999166      0.99572      0.99618
## Pos Pred Value   0.84211      0.91667      0.875000      0.87500      0.86957
## Neg Pred Value   0.99748      0.98292      0.998749      0.99657      0.99915
## Prevalence       0.01573      0.06625      0.007039      0.03230      0.02567
## Detection Rate   0.01325      0.05010      0.005797      0.02899      0.02484
## Detection Prevalence 0.01573      0.05466      0.006625      0.03313      0.02857
## Balanced Accuracy 0.91979      0.87569      0.911348      0.94658      0.98196
##
##      Class: LAML Class: LGG Class: LIHC Class: LUAD Class: LUSC
## Sensitivity      1.00000      0.96875      0.89796      0.87402      0.82143
## Specificity      0.99831      0.99650      0.99741      0.99825      0.99013
## Pos Pred Value   0.91667      0.93939      0.93617      0.96522      0.75000
## Neg Pred Value   1.00000      0.99825      0.99569      0.99304      0.99354
## Prevalence       0.01822      0.05300      0.04058      0.05259      0.03478
## Detection Rate   0.01822      0.05135      0.03644      0.04596      0.02857
## Detection Prevalence 0.01988      0.05466      0.03892      0.04762      0.03810
## Balanced Accuracy 0.99916      0.98263      0.94768      0.93613      0.90578
##
##      Class: MESO Class: OV Class: PAAD Class: PCPG Class: PRAD
## Sensitivity      0.952381 1.0000000      0.95122      0.93617      0.96124
## Specificity      0.999582 0.9995857      0.99705      0.99916      1.00000
## Pos Pred Value   0.952381 0.5000000      0.84783      0.95652      1.00000
## Neg Pred Value   0.999582 1.0000000      0.99916      0.99873      0.99782
## Prevalence       0.008696 0.0004141      0.01698      0.01946      0.05342
## Detection Rate   0.008282 0.0004141      0.01615      0.01822      0.05135
## Detection Prevalence 0.008696 0.0008282      0.01905      0.01905      0.05135
## Balanced Accuracy 0.975982 0.9997929      0.97414      0.96766      0.98062

```

```
##          Class: READ Class: SARC Class: SKCM Class: STAD
## Sensitivity          NA      0.83784    0.80142    0.80000
## Specificity          0.990062    0.99829    0.99780    0.98712
## Pos Pred Value       NA      0.93939    0.95763    0.69388
## Neg Pred Value       NA      0.99489    0.98781    0.99266
## Prevalence           0.000000    0.03064    0.05839    0.03520
## Detection Rate       0.000000    0.02567    0.04679    0.02816
## Detection Prevalence 0.009938    0.02733    0.04886    0.04058
## Balanced Accuracy     NA      0.91806    0.89961    0.89356
##          Class: TGCT Class: THCA Class: THYM Class: UCEC Class: UCS
## Sensitivity          0.87179    1.00000    1.00000    0.95327    0.800000
## Specificity          1.00000    0.99956    0.99832    0.99740    0.999167
## Pos Pred Value       1.00000    0.99213    0.87097    0.94444    0.857143
## Neg Pred Value       0.99790    1.00000    1.00000    0.99783    0.998751
## Prevalence           0.01615    0.05217    0.01118    0.04431    0.006211
## Detection Rate       0.01408    0.05217    0.01118    0.04224    0.004969
## Detection Prevalence 0.01408    0.05259    0.01284    0.04472    0.005797
## Balanced Accuracy     0.93590    0.99978    0.99916    0.97534    0.899583
##          Class: UVM
## Sensitivity          0.937500
## Specificity          0.997916
## Pos Pred Value       0.750000
## Neg Pred Value       0.999582
## Prevalence           0.006625
## Detection Rate       0.006211
## Detection Prevalence 0.008282
## Balanced Accuracy     0.967708
```

1.10 Algoritmo randomforest

```
be <- as.data.frame(betas_train)
be$label <- fenotipos_train
modelo_rf <- randomForest(label ~., data=be, ntree=40,
                           importance=TRUE)

resultado <- predict(modelo_rf, newdata=betas_test, type="class")

c4 <- confusionMatrix(fenotipos_test, resultado)
c4
```

Confusion Matrix and Statistics

```
##
##          Reference
## Prediction ACC BLCA BRCA CESC CHOL COAD Control DLBC ESCA GBM HNSC KICH KIRC
## ACC      16    0    1    0    0    0    0    0    0    0    0    0    0
## BLCA      0   92    1    0    0    0    2    1    0    0    0    0    0
## BRCA      0    0  190    0    0    0    3    0    0    0    0    0    0
## CESC      0    0    0   63    0    0    2    0    1    0    5    0    0
## CHOL      0    0    2    0    4    1    1    0    0    0    0    0    0
## COAD      0    0    0    0    0    71   0    0    0    0    0    0    0
## Control   0    0    2    0    0    0   174   0    0    0    1    0    0
## DLBC      0    0    0    0    0    0    0   10    0    0    0    0    0
```

##	ESCA	0	1	1	2	0	0	0	0	10	0	16	0	0	
##	GBM	0	0	0	0	0	0	0	0	0	33	0	0	0	
##	HNSC	0	0	0	4	0	0	1	0	0	0	119	0	0	
##	KICH	0	1	1	0	0	0	0	0	0	0	0	14	0	
##	KIRC	0	0	0	0	0	0	0	0	0	0	0	0	76	
##	KIRP	0	2	0	0	0	0	0	0	0	0	0	0	3	
##	LAML	0	0	0	0	0	0	0	0	0	0	0	0	0	
##	LGG	0	0	0	0	0	0	1	0	0	1	0	0	0	
##	LIHC	0	1	0	0	2	0	3	0	0	0	0	0	1	
##	LUAD	0	1	0	0	0	0	2	0	0	0	0	0	0	
##	LUSC	0	3	2	2	0	0	1	0	0	0	8	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	1	0	0	3	4	0	0	0	0	0	1	
##	PCPG	0	0	0	0	0	0	1	0	0	0	0	0	0	
##	PRAD	0	0	1	0	0	0	2	0	0	0	0	0	0	
##	READ	0	0	0	0	0	17	0	0	0	0	0	0	0	
##	SARC	0	0	0	0	0	0	0	0	0	0	0	0	0	
##	SKCM	1	0	0	0	0	0	1	0	0	0	0	0	0	
##	STAD	0	2	0	0	0	1	4	1	2	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	1	0	0	0	0	0	0	
##	THYM	0	1	0	1	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	1	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	1	0	0	1	0	0	0	0	0	0	0	0	1	0
##	BLCA	0	0	0	0	0	3	0	0	0	0	0	0	1	2
##	BRCA	0	0	0	0	0	1	0	0	0	0	0	0	2	0
##	CESC	0	0	0	0	0	2	0	0	0	0	0	0	0	0
##	CHOL	0	0	0	1	0	0	0	0	0	0	0	0	0	0
##	COAD	0	0	0	0	0	0	0	0	1	0	0	0	0	0
##	Control	0	0	0	0	0	1	0	0	0	0	4	0	0	0
##	DLBC	0	0	0	1	0	0	0	0	0	0	0	0	0	1
##	ESCA	0	0	0	0	1	3	0	0	0	0	0	0	0	0
##	GBM	0	0	4	0	0	0	0	0	0	0	0	0	1	0
##	HNSC	0	0	0	0	0	8	0	0	0	0	0	0	0	0
##	KICH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	KIRC	0	0	0	0	0	0	1	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	130	0	0	0	0	0	0	0	0	0	0	0
##	LIHC	0	0	0	86	0	0	0	0	0	0	0	0	1	0
##	LUAD	0	0	0	0	110	1	0	0	0	0	0	0	1	0
##	LUSC	0	0	0	0	0	74	0	0	0	0	0	0	0	2
##	MESO	0	0	0	0	0	0	19	0	0	0	0	0	2	0
##	OV	0	0	0	0	0	0	0	1	0	0	0	0	0	0
##	PAAD	0	0	0	1	0	0	0	0	30	0	0	0	1	0
##	PCPG	0	0	0	0	0	0	0	0	0	45	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	5	0	0
##	SARC	0	0	0	0	0	0	1	0	0	0	0	0	63	1

##	SKCM	0	0	0	0	0	0	0	0	0	0	0	0	1	115
##	STAD	0	0	0	0	0	0	0	0	3	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	1	0	0	0	0	0	0	0	0	0	0
##	UCEC	0	0	0	0	0	1	0	0	0	0	0	0	0	1
##	UCS	0	0	0	0	0	0	0	0	0	0	0	0	1	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	1	0	0	0	0	0	0
##	BRCA	1	0	0	0	0	0	0
##	CESC	0	0	0	0	4	0	0
##	CHOL	0	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	12	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	0	0	0
##	LUSC	0	0	0	0	0	0	0
##	MESO	0	0	0	0	0	0	0
##	OV	0	0	0	0	1	0	0
##	PAAD	4	0	0	0	0	0	0
##	PCPG	0	0	0	0	0	0	0
##	PRAD	0	0	0	0	0	0	0
##	READ	2	0	0	0	0	0	0
##	SARC	0	0	0	0	1	0	0
##	SKCM	0	0	0	0	0	0	0
##	STAD	83	1	0	0	1	0	0
##	TGCT	0	34	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	106	0	0
##	UCS	0	0	0	0	7	5	0
##	UVM	0	0	0	0	0	0	20

##

Overall Statistics

##

Accuracy : 0.9048

95% CI : (0.8924, 0.9162)

No Information Rate : 0.0841

P-Value [Acc > NIR] : < 2.2e-16

##

Kappa : 0.9002

##


```

## McNemar's Test P-Value : NA
##
## Statistics by Class:
##
##          Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL
## Sensitivity      0.941176    0.86792    0.94059    0.87500    0.666667
## Specificity      0.998332    0.99524    0.99684    0.99402    0.997924
## Pos Pred Value   0.800000    0.89320    0.96447    0.81818    0.444444
## Neg Pred Value   0.999582    0.99394    0.99459    0.99615    0.999169
## Prevalence       0.007039    0.04389    0.08364    0.02981    0.002484
## Detection Rate   0.006625    0.03810    0.07867    0.02609    0.001656
## Detection Prevalence 0.008282    0.04265    0.08157    0.03188    0.003727
## Balanced Accuracy 0.969754    0.93158    0.96872    0.93451    0.832296
##
##          Class: COAD Class: Control Class: DLBC Class: ESCA
## Sensitivity      0.76344    0.85714    0.833333    0.769231
## Specificity      0.99914    0.99593    0.999168    0.985012
## Pos Pred Value   0.97260    0.95082    0.833333    0.217391
## Neg Pred Value   0.99061    0.98701    0.999168    0.998734
## Prevalence       0.03851    0.08406    0.004969    0.005383
## Detection Rate   0.02940    0.07205    0.004141    0.004141
## Detection Prevalence 0.03023    0.07578    0.004969    0.019048
## Balanced Accuracy 0.88129    0.92654    0.916251    0.877122
##
##          Class: GBM Class: HNSC Class: KICH Class: KIRC Class: KIRP
## Sensitivity      0.97059    0.79866    1.000000    0.93827    0.98462
## Specificity      0.99790    0.99426    0.999167    0.99829    0.99787
## Pos Pred Value   0.86842    0.90152    0.875000    0.95000    0.92754
## Neg Pred Value   0.99958    0.98686    1.000000    0.99786    0.99957
## Prevalence       0.01408    0.06170    0.005797    0.03354    0.02692
## Detection Rate   0.01366    0.04928    0.005797    0.03147    0.02650
## Detection Prevalence 0.01573    0.05466    0.006625    0.03313    0.02857
## Balanced Accuracy 0.98424    0.89646    0.999584    0.96828    0.99124
##
##          Class: LAML Class: LGG Class: LIHC Class: LUAD Class: LUSC
## Sensitivity      1.00000    0.97015    0.94505    0.99099    0.78723
## Specificity      1.00000    0.99912    0.99656    0.99783    0.99224
## Pos Pred Value   1.00000    0.98485    0.91489    0.95652    0.80435
## Neg Pred Value   1.00000    0.99825    0.99785    0.99957    0.99139
## Prevalence       0.01988    0.05549    0.03768    0.04596    0.03892
## Detection Rate   0.01988    0.05383    0.03561    0.04555    0.03064
## Detection Prevalence 0.01988    0.05466    0.03892    0.04762    0.03810
## Balanced Accuracy 1.00000    0.98464    0.97081    0.99441    0.88974
##
##          Class: MESO Class: OV Class: PAAD Class: PCPG Class: PRAD
## Sensitivity      0.904762 1.0000000    0.88235    1.00000    0.96800
## Specificity      0.999165 0.9995857    0.99328    0.99958    0.99869
## Pos Pred Value   0.904762 0.5000000    0.65217    0.97826    0.97581
## Neg Pred Value   0.999165 1.0000000    0.99831    1.00000    0.99825
## Prevalence       0.008696 0.0004141    0.01408    0.01863    0.05176
## Detection Rate   0.007867 0.0004141    0.01242    0.01863    0.05010
## Detection Prevalence 0.008696 0.0008282    0.01905    0.01905    0.05135
## Balanced Accuracy 0.951963 0.9997929    0.93782    0.99979    0.98334
##
##          Class: READ Class: SARC Class: SKCM Class: STAD
## Sensitivity      1.000000    0.80769    0.94262    0.79808
## Specificity      0.992116    0.99872    0.99869    0.99351
## Pos Pred Value   0.208333    0.95455    0.97458    0.84694
## Neg Pred Value   1.000000    0.99361    0.99695    0.99094

```

## Prevalence	0.002070	0.03230	0.05052	0.04306	
## Detection Rate	0.002070	0.02609	0.04762	0.03437	
## Detection Prevalence	0.009938	0.02733	0.04886	0.04058	
## Balanced Accuracy	0.996058	0.90320	0.97066	0.89579	
##	Class: TGCT	Class: THCA	Class: THYM	Class: UCEC	Class: UCS
## Sensitivity	0.97143	0.99213	1.00000	0.88333	1.000000
## Specificity	1.00000	0.99956	0.99874	0.99913	0.996266
## Pos Pred Value	1.00000	0.99213	0.90323	0.98148	0.357143
## Neg Pred Value	0.99958	0.99956	1.00000	0.99393	1.000000
## Prevalence	0.01449	0.05259	0.01159	0.04969	0.002070
## Detection Rate	0.01408	0.05217	0.01159	0.04389	0.002070
## Detection Prevalence	0.01408	0.05259	0.01284	0.04472	0.005797
## Balanced Accuracy	0.98571	0.99584	0.99937	0.94123	0.998133
##	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				

El modelo resultante da un valor de *accuracy* de 0.9 y del parámetro Kappa de 0.9. El valor de la *sensitivity* del modelo es 0.94.

1.11 Best model randomforest con caret

```
ctrl <- trainControl(method = "repeatedcv",
                     number=5, repeats=5,
                     selectionFunction="best",
                     savePredictions=TRUE,
                     classProbs=TRUE,
                     verboseIter=TRUE,
                     allowParallel=TRUE)

grid_rf <- expand.grid(mtry = c(60, 80, 100, 120, 140))

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

```
## Aggregating results
## Selecting tuning parameters
## Fitting mtry = 140 on full training set
```

```
m_rf
```

```
## Random Forest
```

```
##
## 7292 samples
## 1492 predictors
## 34 classes: 'ACC', 'BLCA', 'BRCA', 'CESC', 'CHOL', 'COAD', 'Control', 'DLBC', 'ESCA', 'GBM', 'HNSC'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold, repeated 5 times)
## Summary of sample sizes: 5834, 5832, 5833, 5833, 5836, 5833, ...
## Resampling results across tuning parameters:
##
## mtry Accuracy Kappa
## 60 0.9126427 0.9084980
## 80 0.9145901 0.9105449
## 100 0.9152210 0.9112100
## 120 0.9162094 0.9122483
## 140 0.9177726 0.9138876
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 140.
```

```
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   1   0   0   0   0   0   0   0   0   0   0
## BLCA      0  93   2   0   0   0   0   1   1   0   0   1   0   0
## BRCA      0   0 191   0   0   0   0   3   0   0   0   1   0   0
## CESC      0   0   0  65   0   0   0   2   0   0   0   6   0   0
## CHOL      0   0   1   0   4   0   0   1   0   0   0   0   0   0
## COAD      0   0   0   0   0   72   0   0   0   0   0   0   0   0
## Control   0   0   2   0   0   0   174   0   0   0   0   0   0   0
## DLBC      0   0   0   0   0   0   0   0  11   0   0   0   0   0
## ESCA      0   2   0   1   0   0   0   1   0  14   0  13   0   0
## GBM      0   0   0   0   0   0   0   0   0   0  33   0   0   0
## HNSC      0   2   0   2   0   0   0   1   0   0   0  119   0   0
## KICH      0   1   1   0   0   0   0   0   0   0   0   0  14   0
## KIRC      0   0   0   0   0   0   0   0   0   0   0   0   0  76
## KIRP      0   1   2   0   0   0   0   1   0   0   0   0   0   2
## LAML      0   0   1   0   0   0   0   0   0   0   0   0   0   0
## LGG      0   0   0   0   0   0   0   0   0   0   1   0   0   1
## LIHC      0   0   0   0   1   0   0   2   0   0   0   0   0   0
## LUAD      0   0   0   0   0   0   0   1   0   0   0   0   0   0
## LUSC      0   3   1   1   0   0   0   1   0   0   0   2   0   0
## MESO      0   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   0
## PAAD      0   1   0   0   0   0   0   4   0   0   0   0   0   0
## PCPG      0   0   0   0   0   0   0   1   0   0   0   0   0   0
## PRAD      0   0   0   0   0   0   0   2   0   0   0   0   0   0
## READ      0   0   0   0   0   0  16   0   0   0   0   0   0   0
```

##	SARC	0	0	1	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	0
##	STAD	0	0	0	0	0	1	2	1	3	0	1	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	1	0	0	0	0	0	0	0
##	THYM	0	0	0	0	0	0	0	0	0	0	1	0	0	0
##	UCEC	0	1	0	0	0	0	0	0	0	0	0	0	0	0
##	UCS	0	1	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	KIRP	LAML	LGG	LIHC	LUAD	LUSC	MESO	OV	PAAD	PCPG	PRAD	READ	SARC	SKCM
##	ACC	0	0	0	0	0	0	0	0	0	0	0	0	0	1
##	BLCA	0	0	0	0	0	3	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	1	0	0	0	0	0	0	0	0	0	0
##	COAD	0	0	0	1	0	0	0	0	0	0	0	0	0	0
##	Control	0	0	0	0	0	0	0	0	0	0	4	0	0	0
##	DLBC	0	0	0	0	0	0	0	0	0	0	0	0	0	1
##	ESCA	0	0	0	0	1	4	0	0	0	0	0	0	0	0
##	GBM	0	0	4	0	0	0	0	0	0	0	0	0	1	0
##	HNSC	0	0	0	0	0	7	0	0	0	0	0	0	0	0
##	KICH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
##	KIRC	1	0	0	0	0	0	0	0	0	0	0	0	3	0
##	KIRP	63	0	0	0	0	0	0	0	0	0	0	0	0	0
##	LAML	0	47	0	0	0	0	0	0	0	0	0	0	0	0
##	LGG	0	0	130	0	0	0	0	0	0	0	0	0	0	0
##	LIHC	0	0	0	90	0	0	0	0	0	0	0	0	1	0
##	LUAD	0	0	0	0	113	1	0	0	0	0	0	0	0	0
##	LUSC	1	0	0	0	0	82	0	0	0	0	0	0	0	1
##	MESO	0	0	0	0	0	0	20	0	0	0	0	0	1	0
##	OV	0	0	0	0	0	0	0	1	0	0	0	0	0	0
##	PAAD	0	0	0	1	0	1	0	0	36	0	0	0	0	0
##	PCPG	0	0	0	0	0	0	0	0	0	45	0	0	0	0
##	PRAD	0	0	0	0	0	0	0	0	0	0	122	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	61	3
##	SKCM	0	0	0	0	0	0	0	0	0	0	0	0	0	117
##	STAD	0	0	0	0	1	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	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	0	0	0	0	0							
##	BRCA	0	0	0	0	0	0	0							
##	CESC	1	0	0	0	3	0	0							
##	CHOL	2	0	0	0	0	0	0							
##	COAD	0	0	0	0	0	0	0							
##	Control	0	0	3	0	0	0	0							

##	DLBC	0	0	0	0	0	0	0
##	ESCA	10	0	0	0	0	0	0
##	GBM	0	0	0	0	0	0	0
##	HNSC	1	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	0	0	0
##	LUSC	0	0	0	0	0	0	0
##	MESO	0	0	0	0	0	0	0
##	OV	0	0	0	0	1	0	0
##	PAAD	3	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	1	0	0
##	SKCM	0	0	0	0	0	0	0
##	STAD	87	0	0	0	0	0	0
##	TGCT	0	34	0	0	0	0	0
##	THCA	0	0	126	0	0	0	0
##	THYM	0	0	0	29	1	0	0
##	UCEC	0	0	0	0	107	0	0
##	UCS	0	0	0	0	2	11	0
##	UVM	0	0	0	0	0	0	20

##

Overall Statistics

##

Accuracy : 0.9246

95% CI : (0.9134, 0.9349)

No Information Rate : 0.0836

P-Value [Acc > NIR] : < 2.2e-16

##

Kappa : 0.9211

##

McNemar's Test P-Value : NA

##

Statistics by Class:

##

Class: ACC Class: BLCA Class: BRCA Class: CESC Class: CHOL

Sensitivity 1.000000 0.88571 0.94554 0.92857 0.800000

Specificity 0.999166 0.99567 0.99729 0.99488 0.997925

Pos Pred Value 0.900000 0.90291 0.96954 0.84416 0.444444

Neg Pred Value 1.000000 0.99481 0.99504 0.99786 0.999584

Prevalence 0.007453 0.04348 0.08364 0.02899 0.002070

Detection Rate 0.007453 0.03851 0.07909 0.02692 0.001656

Detection Prevalence 0.008282 0.04265 0.08157 0.03188 0.003727

Balanced Accuracy 0.999583 0.94069 0.97142 0.96173 0.898963

##

Class: COAD Class: Control Class: DLBC Class: ESCA

Sensitivity 0.80899 0.87437 0.846154 0.823529

Specificity 0.99957 0.99594 0.999584 0.986656

Pos Pred Value 0.98630 0.95082 0.916667 0.304348

## Neg Pred Value	0.99274	0.98880	0.999168	0.998734	
## Prevalence	0.03685	0.08240	0.005383	0.007039	
## Detection Rate	0.02981	0.07205	0.004555	0.005797	
## Detection Prevalence	0.03023	0.07578	0.004969	0.019048	
## Balanced Accuracy	0.90428	0.93516	0.922869	0.905092	
##	Class: GBM	Class: HNSC	Class: KICH	Class: KIRC	Class: KIRP
## Sensitivity	0.97059	0.82639	1.000000	0.96203	0.96923
## Specificity	0.99790	0.99428	0.999167	0.99829	0.99745
## Pos Pred Value	0.86842	0.90152	0.875000	0.95000	0.91304
## Neg Pred Value	0.99958	0.98905	1.000000	0.99872	0.99915
## Prevalence	0.01408	0.05963	0.005797	0.03271	0.02692
## Detection Rate	0.01366	0.04928	0.005797	0.03147	0.02609
## Detection Prevalence	0.01573	0.05466	0.006625	0.03313	0.02857
## Balanced Accuracy	0.98424	0.91033	0.999584	0.98016	0.98334
##	Class: LAML	Class: LGG	Class: LIHC	Class: LUAD	Class: LUSC
## Sensitivity	1.00000	0.97015	0.96774	0.98261	0.82828
## Specificity	0.99958	0.99912	0.99828	0.99913	0.99568
## Pos Pred Value	0.97917	0.98485	0.95745	0.98261	0.89130
## Neg Pred Value	1.00000	0.99825	0.99871	0.99913	0.99268
## Prevalence	0.01946	0.05549	0.03851	0.04762	0.04099
## Detection Rate	0.01946	0.05383	0.03727	0.04679	0.03395
## Detection Prevalence	0.01988	0.05466	0.03892	0.04762	0.03810
## Balanced Accuracy	0.99979	0.98464	0.98301	0.99087	0.91198
##	Class: MESO	Class: OV	Class: PAAD	Class: PCPG	Class: PRAD
## Sensitivity	1.000000	1.000000	0.97297	1.00000	0.96825
## Specificity	0.999582	0.9995857	0.99579	0.99958	0.99913
## Pos Pred Value	0.952381	0.5000000	0.78261	0.97826	0.98387
## Neg Pred Value	1.000000	1.0000000	0.99958	1.00000	0.99825
## Prevalence	0.008282	0.0004141	0.01532	0.01863	0.05217
## Detection Rate	0.008282	0.0004141	0.01491	0.01863	0.05052
## Detection Prevalence	0.008696	0.0008282	0.01905	0.01905	0.05135
## Balanced Accuracy	0.999791	0.9997929	0.98438	0.99979	0.98369
##	Class: READ	Class: SARC	Class: SKCM	Class: STAD	
## Sensitivity	1.000000	0.87143	0.94355	0.83654	
## Specificity	0.993353	0.99787	0.99956	0.99524	
## Pos Pred Value	0.333333	0.92424	0.99153	0.88776	
## Neg Pred Value	1.000000	0.99617	0.99695	0.99266	
## Prevalence	0.003313	0.02899	0.05135	0.04306	
## Detection Rate	0.003313	0.02526	0.04845	0.03602	
## Detection Prevalence	0.009938	0.02733	0.04886	0.04058	
## Balanced Accuracy	0.996676	0.93465	0.97156	0.91589	
##	Class: TGCT	Class: THCA	Class: THYM	Class: UCEC	Class: UCS
## Sensitivity	1.00000	0.97674	1.00000	0.93043	1.000000
## Specificity	1.00000	0.99956	0.99916	0.99957	0.998752
## Pos Pred Value	1.00000	0.99213	0.93548	0.99074	0.785714
## Neg Pred Value	1.00000	0.99869	1.00000	0.99653	1.000000
## Prevalence	0.01408	0.05342	0.01201	0.04762	0.004555
## Detection Rate	0.01408	0.05217	0.01201	0.04431	0.004555
## Detection Prevalence	0.01408	0.05259	0.01284	0.04472	0.005797
## Balanced Accuracy	1.00000	0.98815	0.99958	0.96500	0.999376
##	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

Bibliografia

- Capper, David, David TW Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, et al. 2018. «DNA methylation-based classification of central nervous system tumours». *Nature* 555 (7697): 469-74.
- 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.
- Zhou, Wanding, Peter W Laird, y Hui Shen. 2017. «Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes». *Nucleic acids research* 45 (4): e22-22.