

APLICANDO UN MODELO PREDICTIVO DE CLASIFICACIÓN

Sigues avanzando con técnicas de ciencia de datos.

Ahora practicarás con algoritmos de clasificación.

Y calcularás tu primer clasificador.

¡A por ello!

IMPORTA LOS DATOS DE BREAST CANCER

Vamos a crear un clasificador que nos permita cuantificar la probabilidad de tener cáncer maligno o no según las características del análisis de una extracción de tejido.

Para no complicarnos puedes leer la tabla de datos de esta forma:

```
# load libraries
library(mlbench)
library(caret)

# Load data
data(BreastCancer)

validation_index <- createDataPartition(BreastCancer$Class, p=0.80, list=FALSE)
# select 20% of the data for validation
validation <- BreastCancer[-validation_index,]
# use the remaining 80% of data to training and testing the models
dataset <- BreastCancer[validation_index,]
# Remove redundant variable Id
dataset <- dataset[,-1]
# convert input values to numeric
for(i in 1:9) {
  dataset[,i] <- as.numeric(as.character(dataset[,i]))
}
# summary
summary(dataset)
```

Si aplicas este código tendrás unos datos de validación (validation)

Y unos datos para entrenar el modelo que se llaman dataset (training)

El siguiente paso es entrenar un árbol de clasificación (CART) y calcular la confusion matrix.

¡Vamos para allá!

CALCULA EL ÁRBOL DE CLASIFICACIÓN CON LOS DATOS DE TRAINING

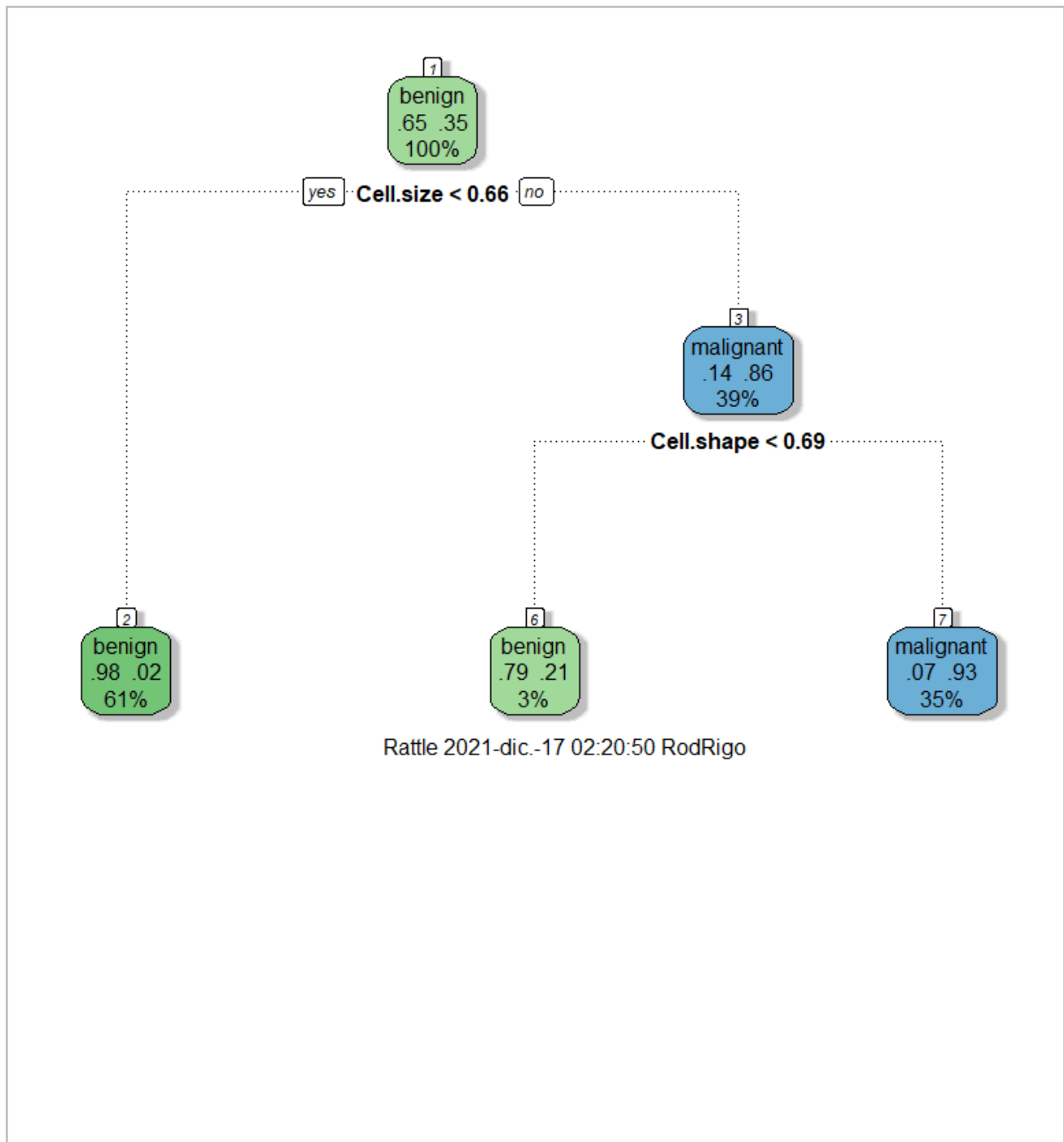
Puedes utilizar la siguiente receta de código para calcular el árbol de clasificación ☺

```
control <- trainControl(method="repeatedcv", number=10, repeats=3)
metric <- "Accuracy"
fit.cart <- train(Class~., data=dataset, method="rpart", metric=metric,
preProc=c("BoxCox"), trControl=control, na.action=na.omit)
```

Dibuja ahora el árbol que acabas de calcular:

```
# plot tree
library(rattle)
fancyRpartPlot(fit.cart $finalModel)
```

Copia el Árbol aquí:



CALCULA LA MATRIZ DE CONFUSIÓN O RESUME EL ACCURACY

Dicho de otro modo, puedes calcular las observaciones que ha acertado el modelo y la que no con la ayuda de los datos de validación.

APLICANDO UN MODELO PREDICTIVO DE CLASIFICACIÓN

Para hacerlo te preparo la receta:

```
# Finalize
data(BreastCancer)
dataset <- BreastCancer
datasetNOMiss <- dataset[complete.cases(dataset),]
# Split out validation dataset
# create a list of 80% of the rows in the original dataset we can use for training
set.seed(7)
validation_index <- createDataPartition(BreastCancer$Class, p=0.80, list=FALSE)
# select 20% of the data for validation
validation <- BreastCancer[-validation_index,]
# use the remaining 80% of data to training and testing the models
dataset <- BreastCancer[validation_index,]
# make predictions
set.seed(7)
x <- datasetNOMiss[,1:9]
preprocessParams <- preProcess(x, method=c("BoxCox"))

predictions <- predict(fit.cart, newdata=preProcess(validation, method=c("BoxCox")))
confusionMatrix(predictions, validation$Class)
```

Copia la matriz de confusión y comenta los resultados
Si estás realmente motivado podrías buscar la manera de dibujar la curva ROC para poder dar un punto más gráfico a tu árbol de clasificación

Confusion Matrix and Statistics

	Reference	
Prediction	benign	malignant
benign	81	5
malignant	6	43

Accuracy : 0.9185
95% CI : (0.8589, 0.9586)
No Information Rate : 0.6444
P-Value [Acc > NIR] : 1.33e-13

Kappa : 0.823

McNemar's Test P-Value : 1

Sensitivity : 0.9310

Specificity : 0.8958

Pos Pred Value : 0.9419

Neg Pred Value : 0.8776

Prevalence : 0.6444

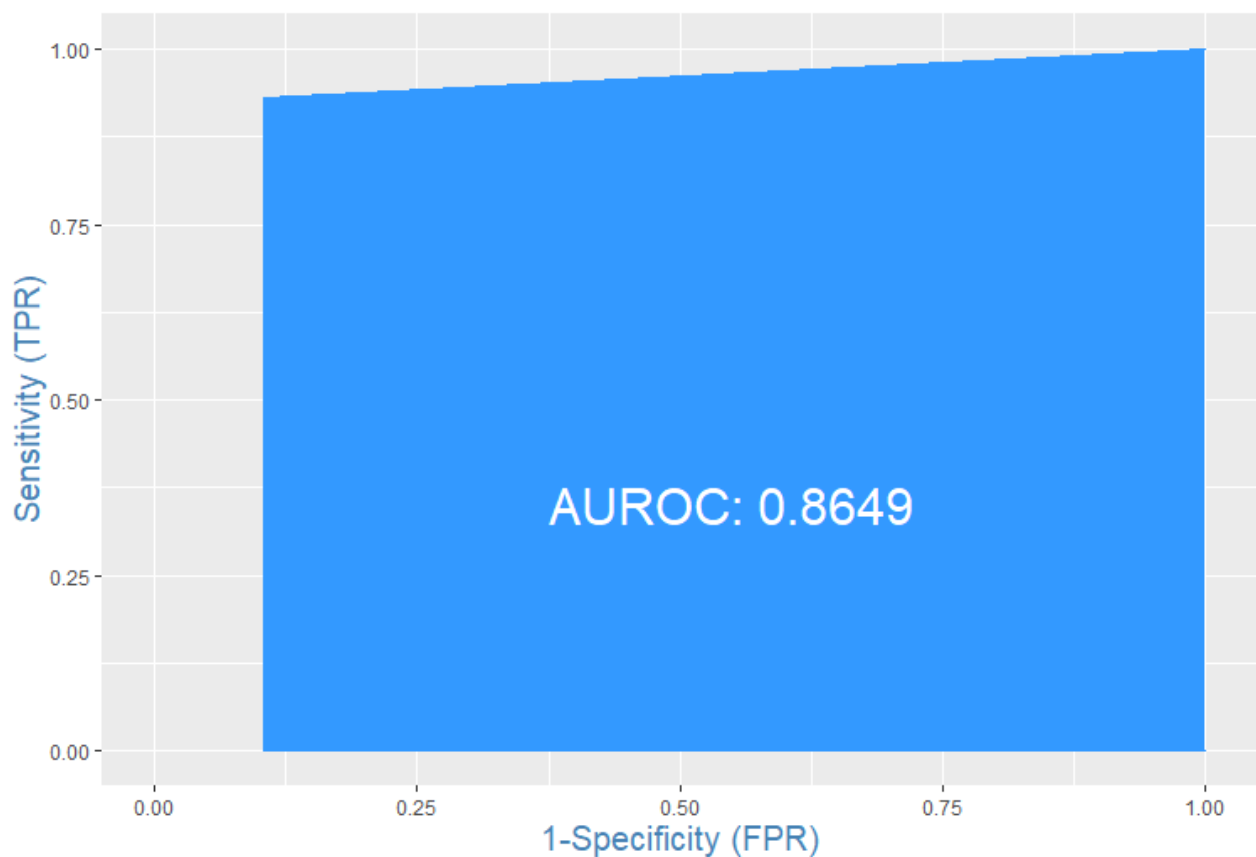
Detection Rate : 0.6000

Detection Prevalence : 0.6370

Balanced Accuracy : 0.9134

'Positive' Class : benign

ROC Curve



No sé si eres consciente pero el trabajo que estás haciendo es increíble.

¡A por esos algoritmos de predicción!

¡Ciao ciao!