Árbol de decisión - Bosque Aleatorio

Efrén Jiménez

1 de setiembre de 2016

Análisis del Problema

En diferentes ramas de la ciencia, como la biología por ejemplo, puede resultar muy importante el poder analizar ágilmente conjuntos de datos de diferentes tamaños con el fin de clasificar especies de plantas, bacterias, animales u otros tipos de organismos. En este caso, se va a intentar crear un modelo que clasifique la especie de diferentes flores basándose en características del sépalo y el pétalo.

Las aplicaciones de dicho modelo pueden ser varias y para diferentes audiencias. Por ejemplo en el contexto de un laboratorio, se puede utilizar para concentrar los esfuerzos en la recolección de muestras y dejar el trabajo de clasificación para el algoritmo. De esta manera, se pueden obtener más muestras en un período menor de tiempo, incluso hasta podría haber un ahorro de dinero significativo al reducir la cantidad de horas necesarias para cumplir con una cuota de muestras.

Entendimiento de los Datos

El conjunto de datos a ser utilizado contiene 150 observaciones, con las siguientes varaibles o columnas:

- Sepal.Length: longitud del sépalo; numérica con valores entre 4.3 y 7.9.
- Sepal.Width: ancho del sépalo; numérica con valores entre 2 y 4.4.
- Petal.Length: largo del pétalo; numérica con valores entre 1 y 6.9.
- Petal.Width: ancho del pétalo; numérica con valores entre 0.1 y 2.5.
- Species: especie a la cual pertenece cada observación; valores posibles: setosa, versicolor y virginica.

Exploración de los Datos

El conjunto de datos que se va a analizar contiene 150 observaciones, 50 de cada especie:

```
#librerías utilizadas
library(caTools)
library(rpart)
library(rpart.plot)
library(rattle)
```

```
## Rattle: A free graphical interface for data mining with R.
## Versión 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Escriba 'rattle()' para agitar, sacudir y  rotar sus datos.
```

```
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

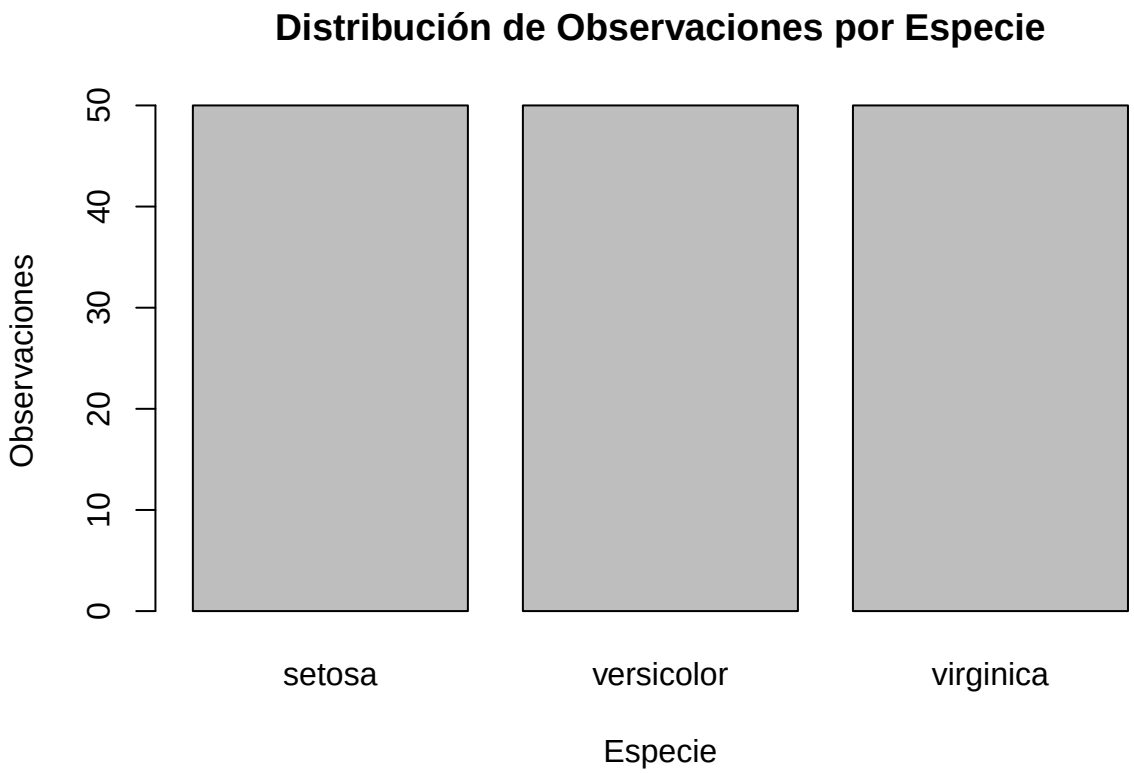
```r
#cargar el conjunto de datos
data("iris")

summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

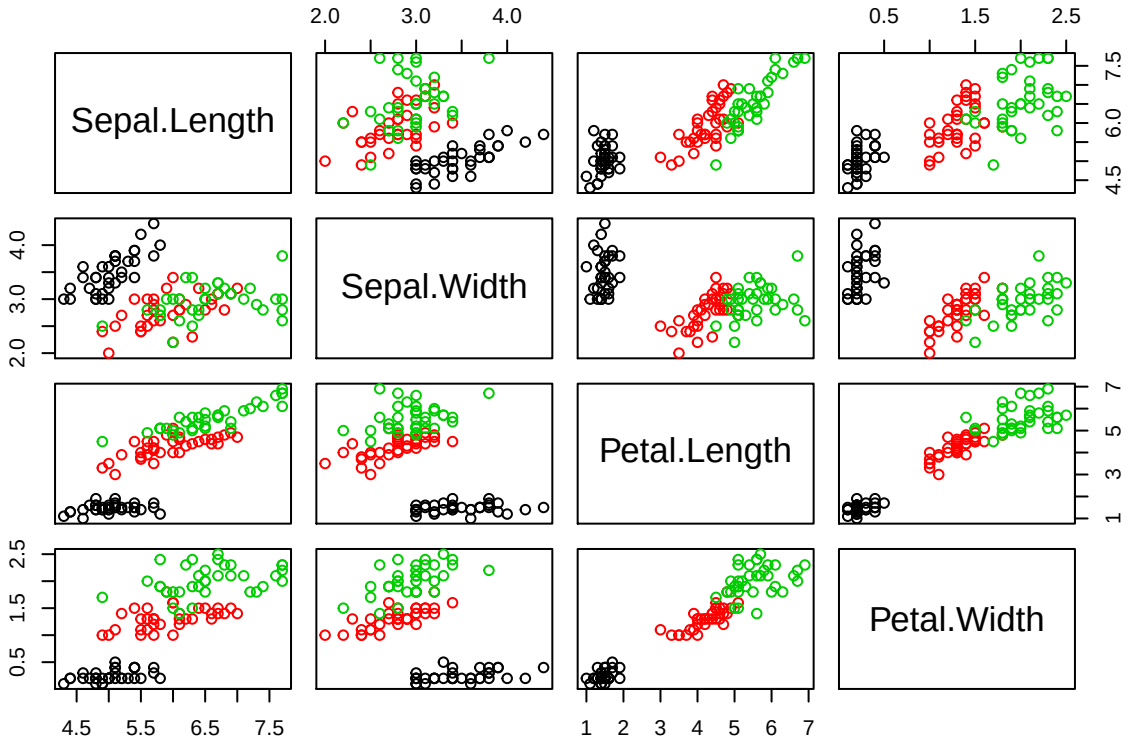
```r
#dividir el conjunto de datos en entrenamiento y prueba
splt <- sample.split(iris$Species, SplitRatio = 0.7)
entrenamiento <- iris[splt,]
prueba <- iris[!splt,]

barplot(table(iris$Species),
        main = 'Distribución de Observaciones por Especie',
        ylab = 'Observaciones',
        xlab = 'Especie')
```



Luego de dividir el conjunto de datos en entrenamiento y prueba, se puede ver como la combinación de diferentes pares de variables muestran una clara división entre cada una de las especies:

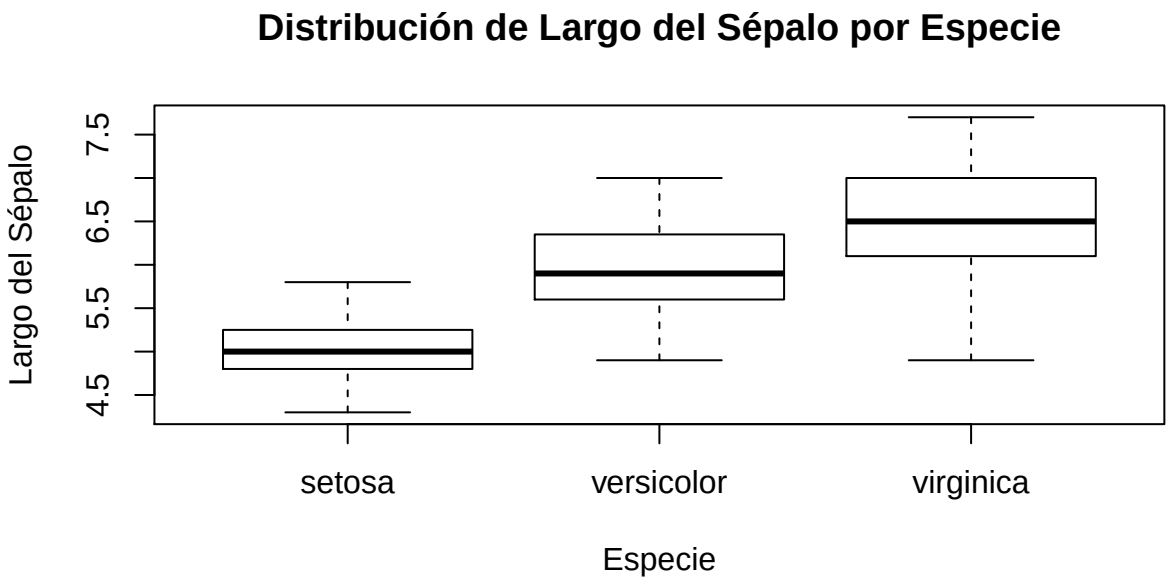
```
pairs(entrenamiento[, -5],
      col = as.numeric(entrenamiento$Species))
```



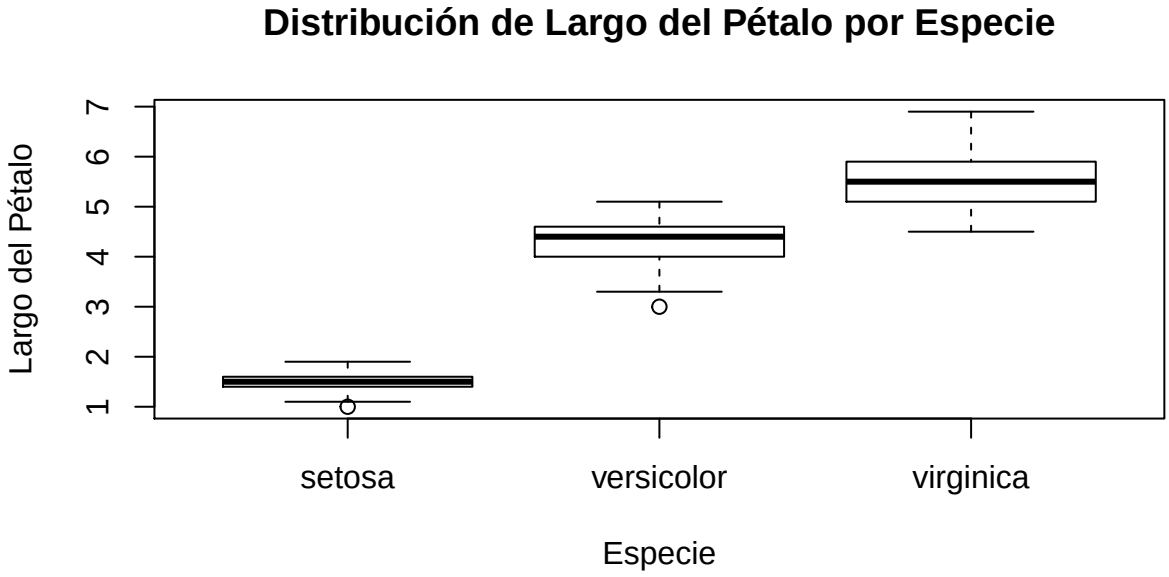
Por ejemplo, la relación entre las variables del ancho y el largo del pétalo permite ver una división considerablemente clara entre las 3 especies. La división no es tan clara cuando se combinan variables como el largo del sépalo y el largo o el ancho del pétalo, pero esas relaciones son bastante más claras que cuando se combinan el largo del sépalo y el ancho del sépalo.

Si se analizan las variables individualmente, se puede apreciar que cada variable aporta información valiosa para la clasificación, pero las que tienen información relacionada con el pétalo son las que presentan las divisiones más claras.

```
boxplot(Sepal.Length ~ Species,
        data = entrenamiento,
        main = 'Distribución de Largo del Sépalo por Especie',
        xlab = 'Especie',
        ylab = 'Largo del Sépalo')
```

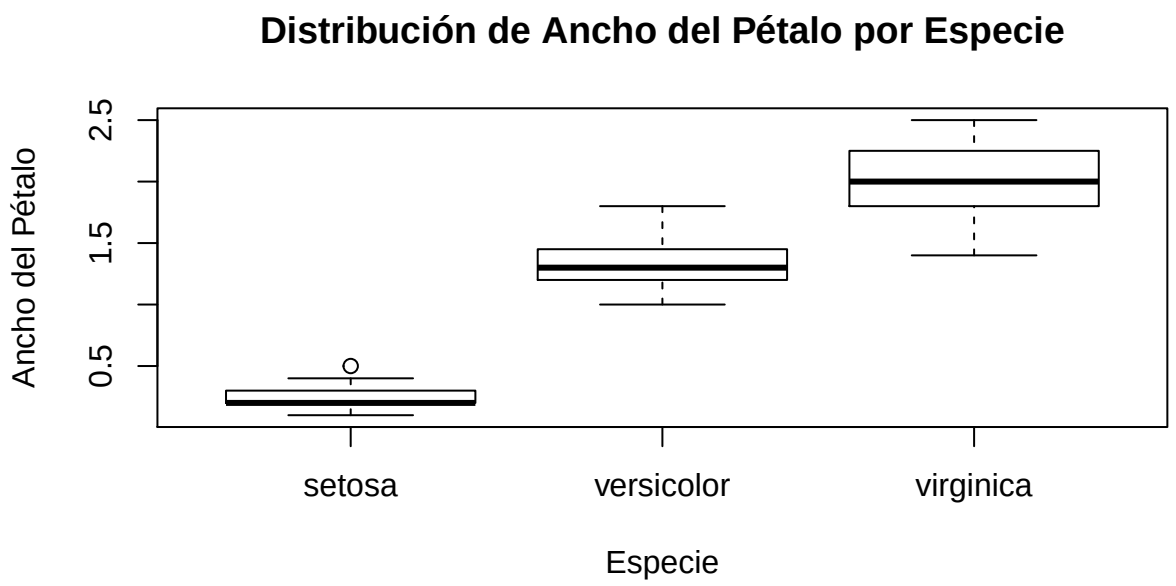



```
boxplot(Petal.Length ~ Species,
        data = entrenamiento,
        main = 'Distribución de Largo del Pétalo por Especie',
        xlab = 'Especie',
        ylab = 'Largo del Pétalo')
```



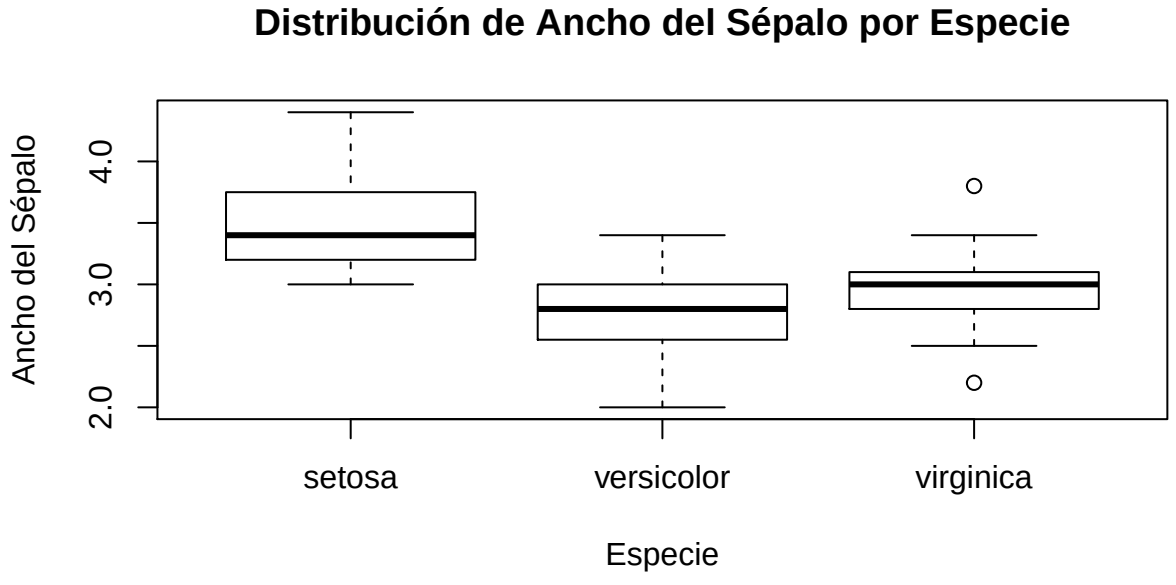

```
boxplot(Petal.Width ~ Species,
        data = entrenamiento,
        main = 'Distribución de Ancho del Pétalo por Especie',
```

```
        xlab = 'Especie',
        ylab = 'Ancho del Pétalo')
```



La variable que presenta divisiones menos claras es la del ancho del sépalo:

```
boxplot(Sepal.Width ~ Species,
        data = entrenamiento,
        main = 'Distribución de Ancho del Sépalo por Especie',
        xlab = 'Especie',
        ylab = 'Ancho del Sépalo')
```

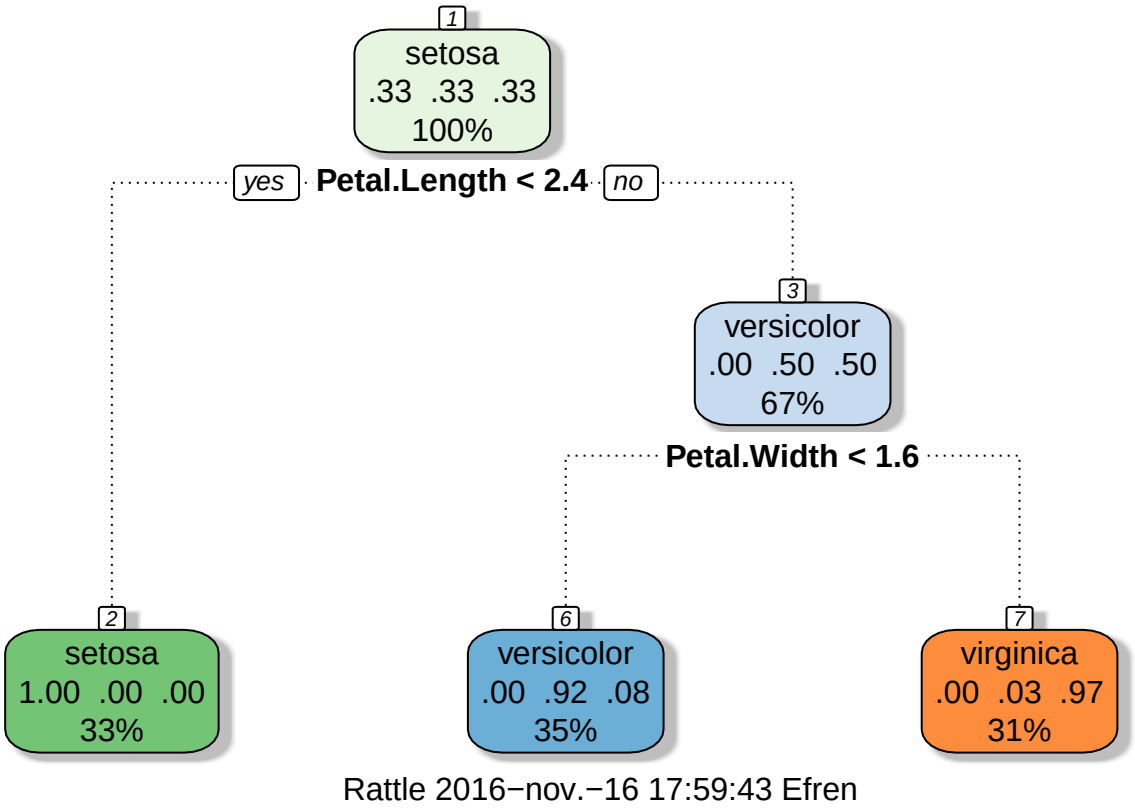


Modelo de Minería de Datos

El primero modelo que se va a utilizar es el de árboles de decisión:

```
modelo.arbol <- rpart(Species ~ .,
                      data = entrenamiento)

fancyRpartPlot(modelo.arbol)
```




```
predicciones.arbol <- predict(modelo.arbol, newdata = prueba, type = 'class')
```

Como se puede apreciar en el gráfico anterior. El modelo determinó que las variables relacionadas con los pétalos son las más importantes para hacer la clasificación.

Alternativamente, se va a crear también un bosque aleatorio:

```
set.seed(4527)
modelo.bosque <- randomForest(Species ~ .,
                              ntrees = 15,
                              data = entrenamiento)

predicciones.bosque <- predict(modelo.bosque, newdata = prueba, type = 'class')
```

Evaluación

Debido a que la variable Especie tiene 3 posibles valores, la evaluación de los modelos se va a centrar en la métrica *exactitud*:

```
table(prueba$Species, predicciones.arbol)
```

```
##             predicciones.arbol
##              setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         14         1
##   virginica       0          1        14
```

El modelo de árbol de decisión clasificó correctamente 43 observaciones de 45, para una exactitud del 95.56%.

```
table(prueba$Species, predicciones.bosque)
```

```
##             predicciones.bosque
##              setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         14         1
##   virginica       0          0        15
```

El bosque aleatorio clasificó correctamente 44 observaciones de 45, para una exactitud del 97.78%. El único error fue una flor virginica clasificada como versicolor.

Resultados

En general, ambos modelos presentan muy buen desempeño, con exactitudes por encima del 95%. Sin embargo el bosque aleatorio tiene una exactitud mayor. Se puede concluir que el caso se presta bastante para un modelo de clasificación, el cual podría ser útil en diferentes escenarios.