Report about the
CREATION OF A CLASSIFICATION TREE
and its implementation on the Iris dataset

GENÍS LÁINEZ MORENO,

Univertitat Autònoma de Barcelona. 8/04/2020

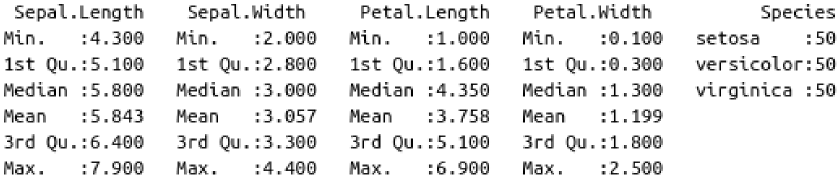
This report is done by answering the **third** exercice of Session 2: **Classification Trees** with *R*

Abstract

This report is intended to make a description of the Iris dataset as well as some plots that could help us to visualize it. Furthermore a classification tree is fitted in order to discriminate different categories (flower species in our case) and its predictive value it has been also calculated.

1 Dataset description and visualization

First a summary it has been done in order to have a first idea of the dataset we are dealing with. This can be seen on the figure 1



```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 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
```

Figure 1: Summary of some characteristics on the Iris Dataset.

One can see that we have four features describing each individual flower. The flowers can be of three types, Setosa, Versicolor and Virginica and the dataset consists on 50 images of each one. Other important data about each feature can also be seen on the figure 1. However, since the goal is going to be the make of a classification tree It could be useful to represent some visual plots that give us the intuition of what are the differences about the species we are dealing with.

Some different plots have been done in order to visualize the data and after this procedure I have concluded that the Sepal Width is the less important of all features in order to distinguish between species. Then, I have plotted the "feature" vs "feature" graph that can be seen on the figure 2. One has to interpret that the text on the squares on the diagonal give the name of the axis adjacent to them.

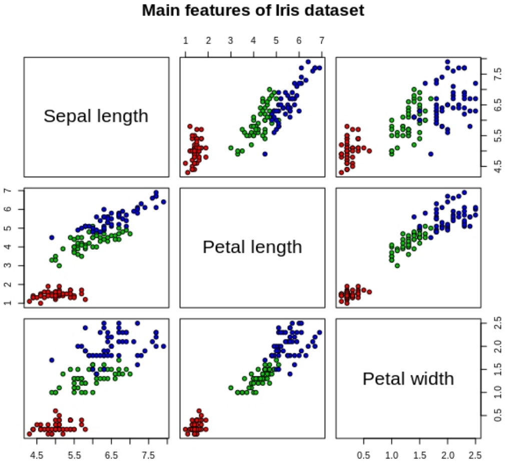


Figure 2: Different plots of the three main characteristics compared. The colors red, green and blue representing the species Setosa, Versicolor and Virginica respectively

We can see that arrived this point is easy to see some different clusters for each spiece. For example, the Setosa one (the red one in the plot) is ”far” from the other two clusters on the ”Petal length” vs. ”Petal With” plots. In fact, what we are instinctively doing is making some first braches of a classification tree. On the first section we are going to create one on the propper way and test its accuracy.

Another interesting plot that gives in fact the very same information (but maybe not as well structured) than the previous one is a 3D plot where the three axis are the chosen features as we can see on the figure 3.

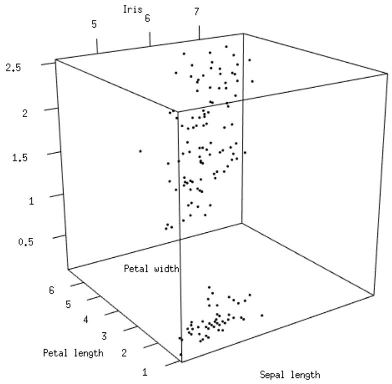


Figure 3: 3d plot of the comparison of the three main characteristics.

2 Classification Tree to discriminate

Arrived this point we can create a classification tree with the help of R code. Since we are going to test the clasification tree a first split of the data has to be done making the Training and the Test datasets (beeing $\frac{2}{3}$ and $\frac{1}{3}$ of the whole dataset respectively). By doing so, We are losing an important part of the dataset to do the tree on the first place since we cannot test on the very same data that we use to train. However, it is a good price to pay If we want to know it advanced how good is our model.

The tree can be seen on the figure 4.

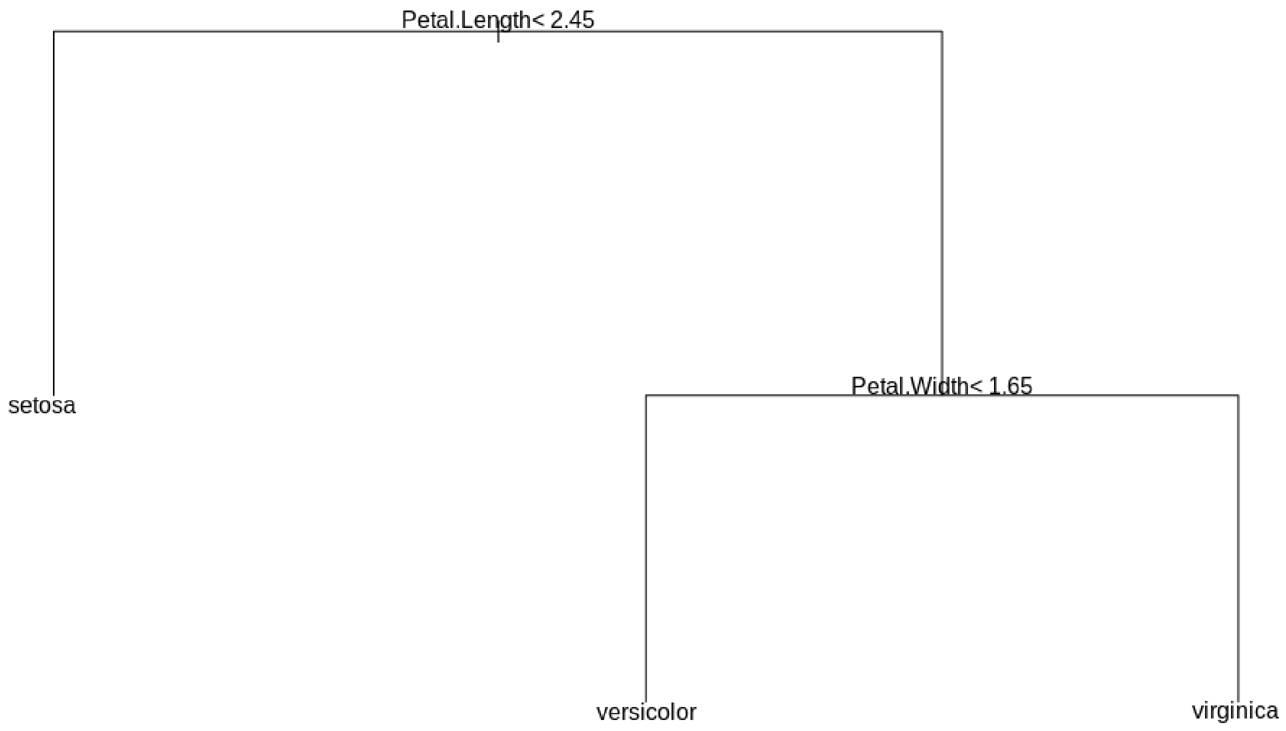


Figure 4: Classification tree on the training subdataset of Iris to make Species classification.

Once the tree has been done, we could make some important comments on it. We just had to use two over the four features that the dataset bring us. Another important comment is that arrived this point one can notice that the tree is consistent with the plots on the figure 2.We could force the tree to go deeper and use other features but as we will see it already gives very good results.

If we apply this classification tree on the train dataset we are going to obtaing the Confusion Matrix for the internal validation process. This matrix can be seen on the figure 5

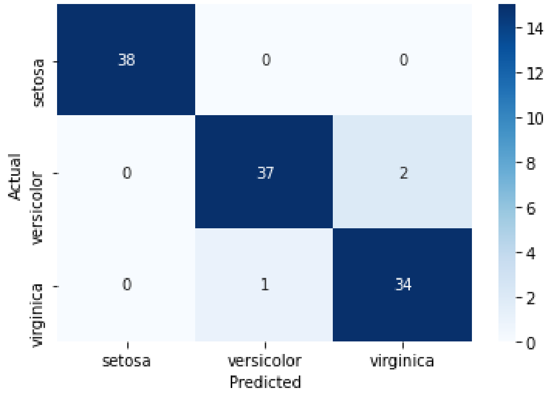


Figure 5: Confusion matrix of the model on the training dataset (internal validation).

We obtain an internal validation accuracy of 97%. This means that the data was pretty distinguishable on the first place like we can see on the figure 2. However, if we want to know the accuracy of the external validation or simply the ”accuracy of the model” we have to apply our tree on the test dataset. This is what one could see on the figure 6.

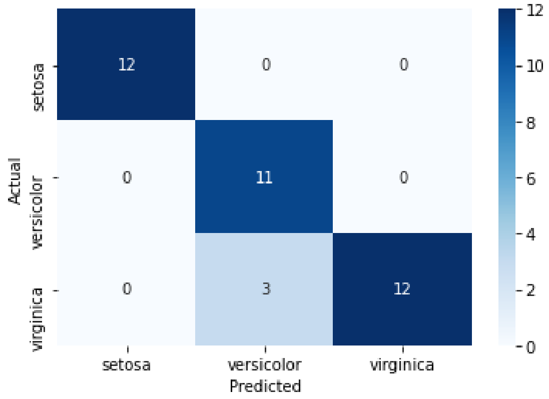


Figure 6: Confusion matrix of the model on the test dataset (external validation)

We can conclude that the model have an accuracy of the 92%. This number is less that the internal validation accuracy as one should expect. We can also see that the confusion matrix is consistent with the results on the previous section since we can see that the Setosa is very well distinguished and the Versicolor and Virginica are more likely to be confused.

3 Further research

Other techniques can be easily applied in order to improve the model. For example using crossvalidation, a deeper study on the percentage we use to train and test the model and the incorporation of other features to force the tree to go deeper on the clasification.

None of this Ideas were intended to be used on this report but all of them are worth to be mentioned.

4 Annex

R Code for the section 1

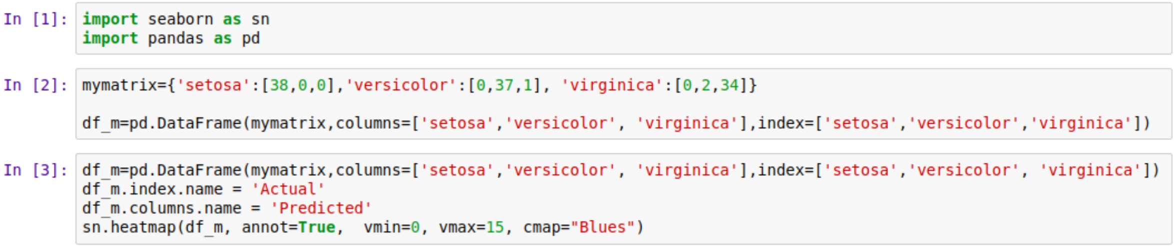
```
#Genis Lainez Moreno

data(iris)
head(iris)

#Summary
summary(iris)

# Consider only the subset of the THREE CHOOSEN FEATURES
data_iris = iris[, c(1,3,4)]
head(data_iris)
colnames(data_iris) <- c("Sepal length", "Petal length", "Petal width")
head(data_iris)

# Visualize the subseted dataset
pairs(data_iris , main = "Main features of Iris dataset", pch = 21, bg = c("red", "green3", "blue")[unclass(iris$Species)])

# 3D Plot 3 main features
X = data_iris
library(rgl)
par(mfrow=c(1,2))
plot3d(X,xlab="Sepal length",ylab="Petal length", zlab = "Petal width",main="Iris")
```

R Code for the section 2

```
#Genis Lainez Moreno

#Data
data(iris)
head(iris)
library(faraway)
iris <- iris[,-2]
library(rpart)
iris.rpart <- rpart(Species~.,data=iris)
iris.rpart
plot(iris.rpart) ; text(iris.rpart)

printcp(iris.rpart)

#Internal and External Validation
dim(iris)
 set.seed(1)
training <- sample(1:150,size=150/4*3)
test <- (1:150)[! (1:150 %in% training)]
sort(training)
sort(test)
iris.training <- iris[training ,]
iris.test <- iris[test ,]

iris.rpart <- rpart(Species~.,data=iris.training)

#internal validation
taula2 <- table(real=iris.training$Species,predit=predict(iris.rpart,type="class"))
taula2

sum(diag(taula2))/sum(taula2)

#external validation
taula3 <- table(real=iris.test$Species,predit=predict(iris.rpart,newdata=iris.test,type="class"))
```

```
taula3
sum(diag(taula3))/sum(taula3)
```

Python Jupyter Notebook for printing the confusion matrices of section 2



```
In [1]: import seaborn as sn
        import pandas as pd

In [2]: mymatrix={'setosa':[38,0,0],'versicolor':[0,37,1], 'virginica':[0,2,34]}

        df_m=pd.DataFrame(mymatrix,columns=['setosa','versicolor', 'virginica'],index=['setosa','versicolor','virginica'])

In [3]: df_m=pd.DataFrame(mymatrix,columns=['setosa','versicolor', 'virginica'],index=['setosa','versicolor', 'virginica'])
        df_m.index.name = 'Actual'
        df_m.columns.name = 'Predicted'
        sn.heatmap(df_m, annot=True,  vmin=0, vmax=15, cmap="Blues")
```