Romane THOIREY
Bixente GRANDJEAN

Project Rapport

Iris Data set

Romane THOIREY
Bixente GRANDJEAN

Table of contents

Romane THOIREY
Bixente GRANDJEAN

INTRODUCTION

In the context of Applied Statistic course, we have to treat about the Iris dataset.

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variable's sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

Iris is a data frame with 150 cases (rows) and 5 variables (columns) named *Sepal.Length, Sepal.Width, Petal.Length, Petal.Width,* and *Species*.

1) DATA EXPLORATION

1) Choose and load the R dataset corresponding to your group subject's and Identify which variables your data set are numeric, and which are categorical (factors) if applicable

We have one categorical value in our data set, the species. The four others are numerical.

2) Generate summary level descriptive statistics: Show the mean, median, 25th and 75th quartiles, min, and max for each of the applicable variables in your data set

```
 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                  
```

3) Determine the frequency for each of one of the categorical variables.
4) Determine the frequency for each of the one of the categorical variables, by a different categorical variable.

There are 50 specimens of each species of iris, which means that the frequency of each one of the categorical variables is of 1/3.

2) GRAPHIC DATA REPRESENTATION

A. UNIVARIATE STUDY

5) Use the commands ***pie() barplot()*** and ***dotchart()*** to represent the categorical data. Comment.

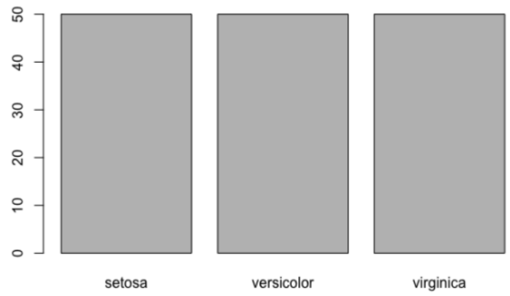

FIGURE 1. BARPLOT OF SPECIES

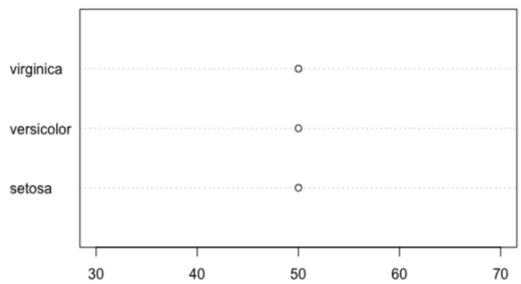

FIGURE 2. DOTCHART OF SPECIES

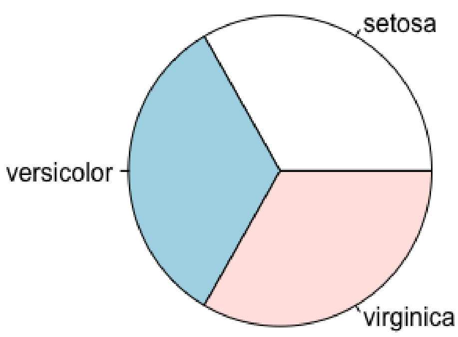

FIGURE 3. PIECHART OF SPECIES

We did the barplot and dotchart for the categorical data but it was not that relevant for our study since we have a good repartition for each categorical data.

6) Create a graph for each single numeric variable. (histogram)

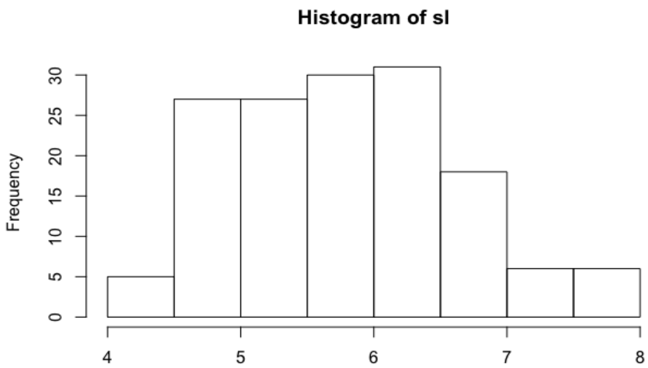

FIGURE 4. HISTOGRAM OF SEPAL LENGTH

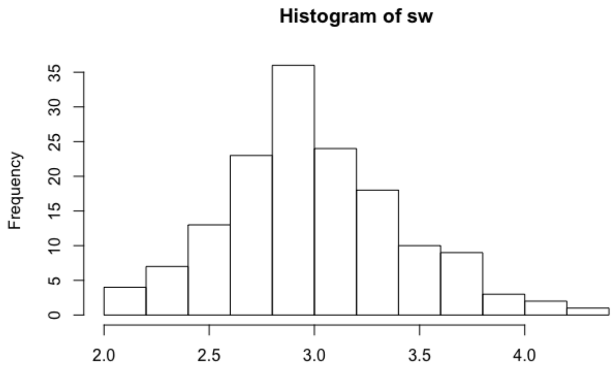

FIGURE 5. HISTOGRAM OF SEPAL WIDTH

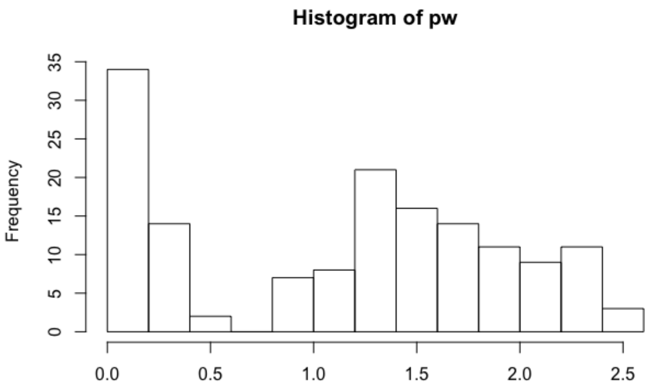

FIGURE 6. HISTOGRAM OF PETAL WIDTH

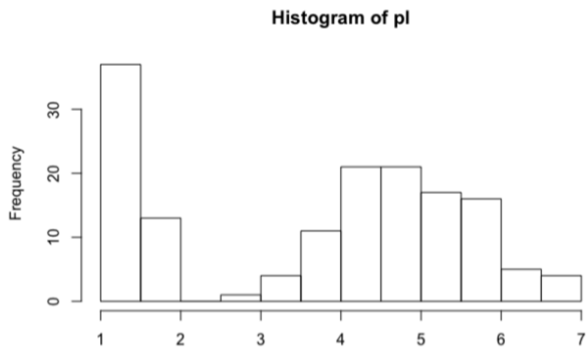

FIGURE 7. HISTOGRAM OF PETAL LENGTH

7) Create a graph for each single numeric corresponding to different categorical variable (histogram)

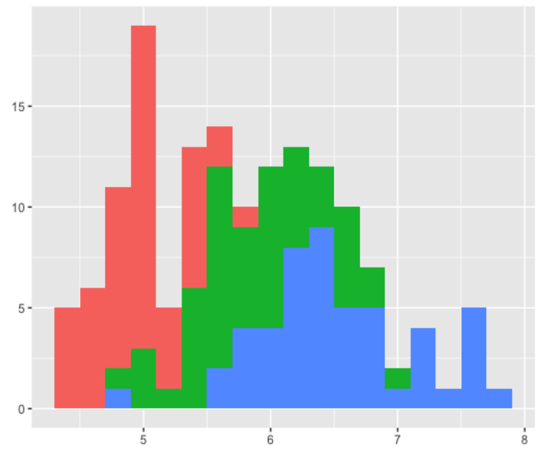

FIGURE 8. PLOT SEPAL LENGTH PER SPECIES

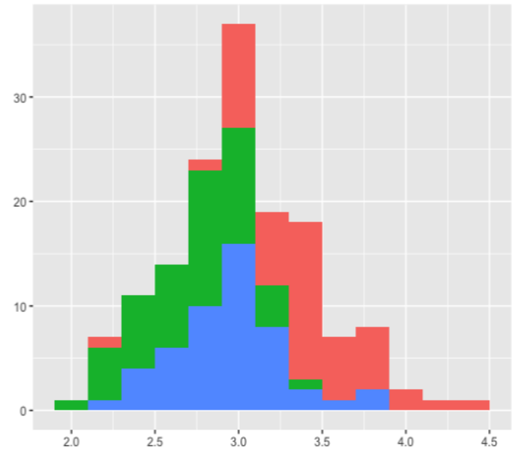

FIGURE 9. PLOT SEPAL WIDTH PER SPECIES

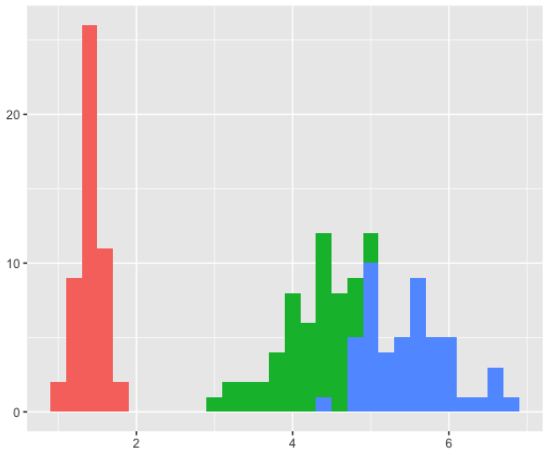

FIGURE 10. PLOT PETAL LENGTH PER SPECIES

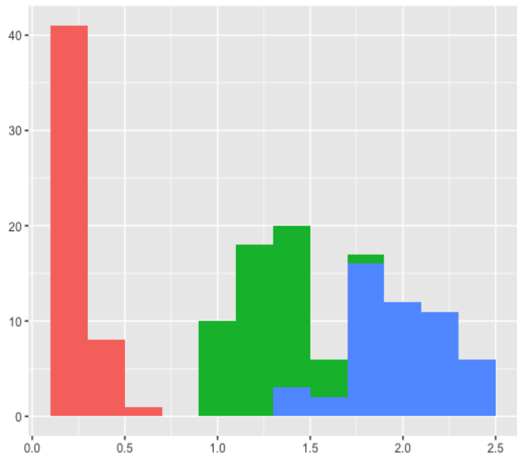

FIGURE 11. PLOT PETAL WIDTH PER SPECIES

B. BIVARIATE STUDY

8) Use the command plot or sunflower plot to plot the scatterplot of the dependent and independent variables. What is the difference between these two commands? Comment your results.

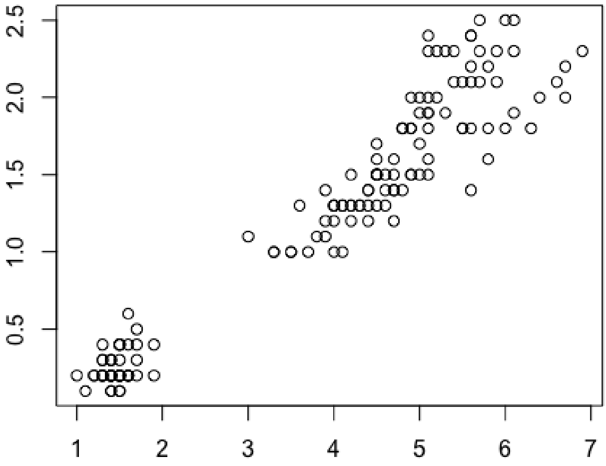

FIGURE 12. PLOT PETAL LENGTH BY WIDTH

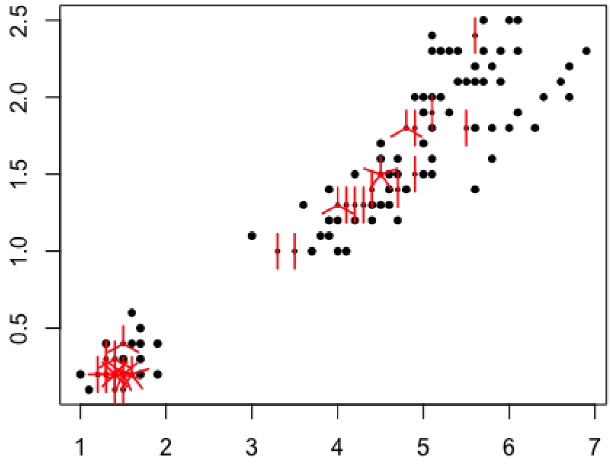

FIGURE 13. SUNFLOWER PLOT PETAL LENGTH BY WIDTH

With the plot we can see that some of observations have been plotted on top of each other. Sunflower plot indicates this number via the petal of the sunflower.

C. Graphic representation for the different data categories

9) Represent the scatter plot for the dependent and independent variables for each data category. Comment.

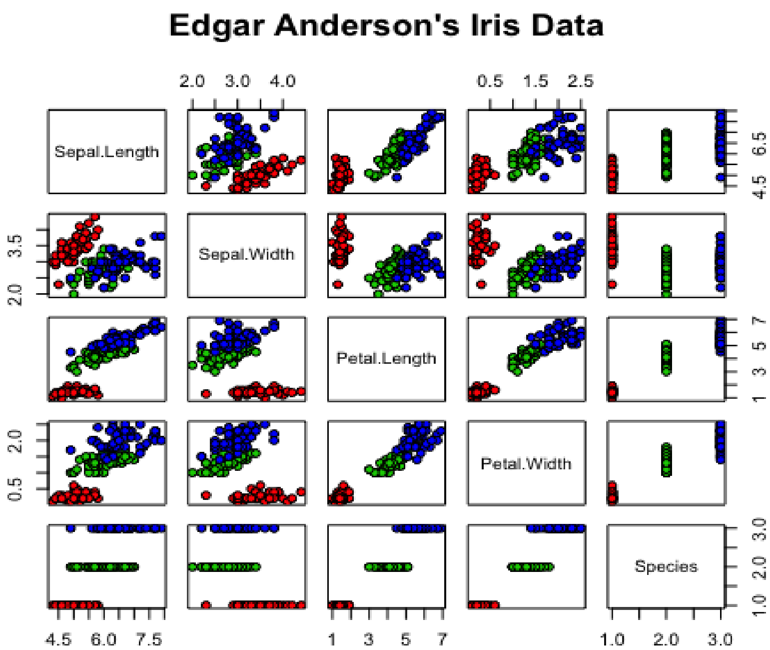

FIGURE 14. PLOT IRIS DATA FOR EACH VARIABLE

Petal.Length and Petal.Width are the most useful features to identify various flower types.

While Setosa can be easily identified (linearly separable, red points), virginica and Versicolor have some overlap (almost linearly separable).

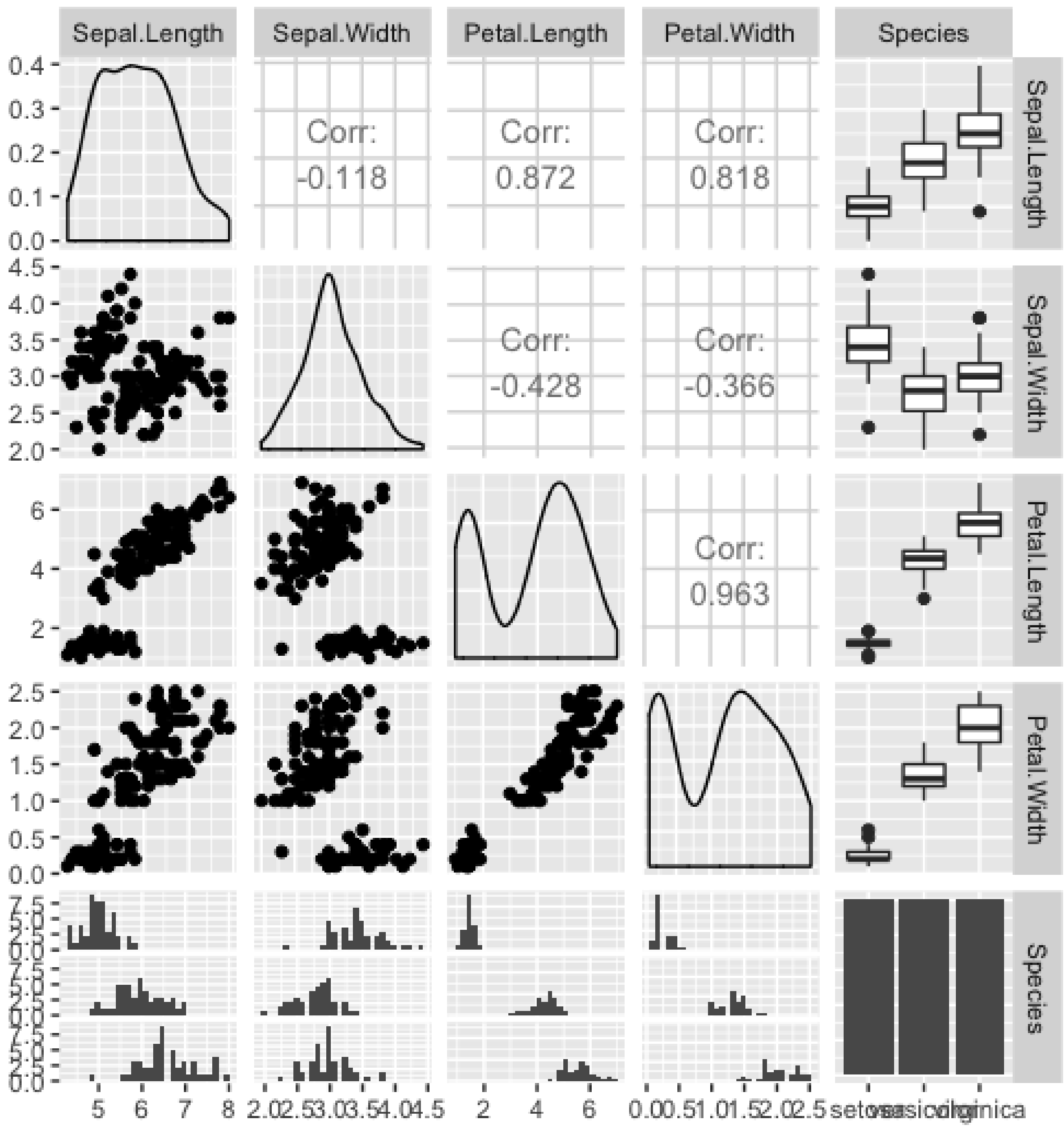

FIGURE 15. PLOT IRIS DATASET

Romane THOIREY
Bixente GRANDJEAN

3) Data prediction and classification

A. Testing hypothesis

10) Use the Chi-square test to verify the independent of each data category.

	Sepal. Length	Sepal.Width	Petal.Length	Petal.Width	Species
Sepal.Length	-	0.1734975	3.764941e-17	3.578992e-02	6.665987e-09
Sepal.Width	-	-	4.221952e-02	1.274332e-04	6.016031e-05
Petal.Length	-	-	-	5.142730e-09	1.177567e-21
Petal.Width	-	-	-	-	2.164810e-35
Species	-	-	-	-	-

As we can see from the result, the p-value is smaller than the threshold value of 5% for each pair of categories except for the Sepal Length and the Sepal Width. This enable us to safely reject the null hypothesis and accept the alternate hypothesis.

B. Build the regression model

- Linear Discriminant Analysis

Our goal is to predict the species of a sample using the features. To build this classification model, we must split data into two sets, the training and the testing sets. For this model, we take an 80/20 repartition: this means that 80% of the samples are in the training set (120 samples) and the remaining 20% are in the testing set (30 samples).

Once we've trained our model on the 120 samples, we can apply it to the test set of 30 values.

```
120 samples
  4 predictor
  3 classes: 'setosa', 'versicolor', 'virginica'
```

Thus, those predictions depend on the training set. Indeed, each time we run the script, train and test data are randomly split so it can lead to different results. Nevertheless, results are, each time, very consistent. We have never had failures for the Setosa but it happens that virginica and versicolor prediction are sometimes wrong.

```
            Reference
Prediction   setosa versicolor virginica
  setosa         10          0         0
  versicolor      0          9         1
  virginica       0          1         9
```

This can be explained by the main difference between Setosa and other species. As we've seen on the figure 14, the red dots (Setosa) are not mixed with the others whereas virginica and versicolor share some common values.

- Classification Tree

The basic idea of a classification tree is to first start with all variables in one group. Then find some characteristic that best separates the groups, for example the first split could be asking whether petal widths are less than or greater than 0.8. Then continue this process until the partitions have sufficiently homogeneous or are too small.

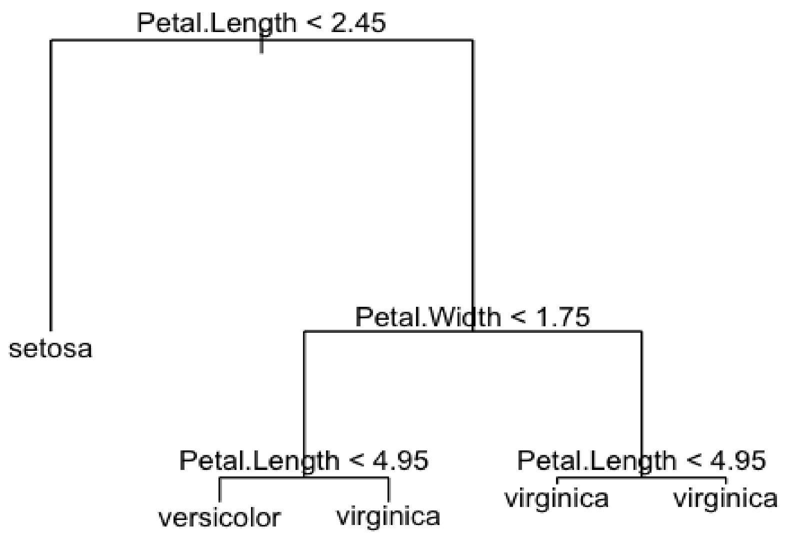

FIGURE 16. CLASSIFICATION TREE USING PETAL LENGTH AND

We used two variables above, Petal.Width and Petal.Length to illustrate the classification process. We can include all four variables in the classification process:

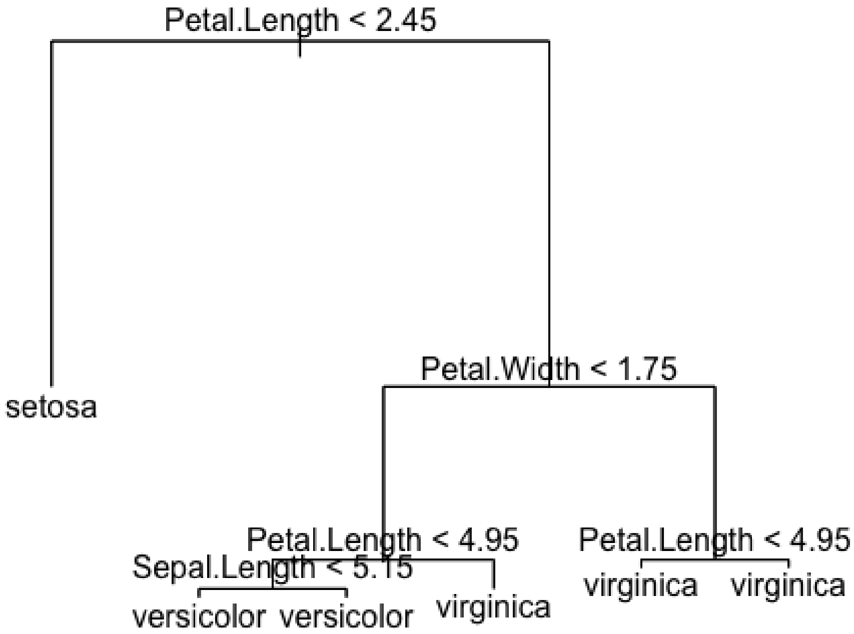

FIGURE 17. CLASSIFICATION TREE OF SPECIES WITH ALL FEATURES

C. Verify model significance (Model validation)

- Linear Discriminant Analysis

	Class: setosa	Class: versicolor	Class: virginica
Sensitivity	1.0000	0.9000	0.9000
Specificity	1.0000	0.9500	0.9500
Pos Pred Value	1.0000	0.9000	0.9000
Neg Pred Value	1.0000	0.9500	0.9500
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.3333	0.3000	0.3000
Detection Prevalence	0.3333	0.3333	0.3333
Balanced Accuracy	1.0000	0.9250	0.9250

Since our model's sensivity (true positive rate) is over 90% for each species (in most cases) – even 100% for the Setosa's – and specify (true negative rate) is over 95%, we can conclude that the predictions are accurate.

- Classification Tree

For the Classification Tree where we used the Petal.Length and the Petal.Width variable, there is only 4 flowers which doesn't fit our model. Our accuracy is up to 97,33% since we have a misclassification rate of 2,67%.

```
Classification tree:
tree(formula = Species ~ Petal.Length + Petal.Width, data = iris)
Number of terminal nodes:  5
Residual mean deviance:  0.157 = 22.77 / 145
Misclassification error rate: 0.02667 = 4 / 150
```

In this case where we use all features and it seems that we have the same accuracy and misclassification rate. We have one more terminal node.

```
Classification tree:
tree(formula = Species ~ Sepal.Width + Sepal.Length + Petal.Length +
    Petal.Width, data = iris)
Variables actually used in tree construction:
[1] "Petal.Length" "Petal.Width"  "Sepal.Length"
Number of terminal nodes:  6
Residual mean deviance:  0.1253 = 18.05 / 144
Misclassification error rate: 0.02667 = 4 / 150
```