Advanced Data Analysis_ESW5024_41 Assignment 1

Jerbi Olfa 2021713094

Used dataset: iris

Source: https://archive.ics.uci.edu/ml/datasets/iris or 'datasets' package on R.

Used program: R

Dataset description: "The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris Setosa, Iris Virginica and Iris Versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms."

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

   -- Iris Setosa

   -- Iris Versicolour

   -- Iris Virginica

Dataset summary:

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	

Using the summary function on R we can see an overview of the whole dataset: the number of datapoints for each iris species (50 each) and for each individual (in this case each iris) the minimum, maximum, median, mean, the 1st and 3rd quantile for each of the four features.

We can also plot the boxplots for each feature to be able to see similar results to the ones shown in summary. However, through box plots it could be easier to compare the distribution for each feature, as well as observing the outliers:

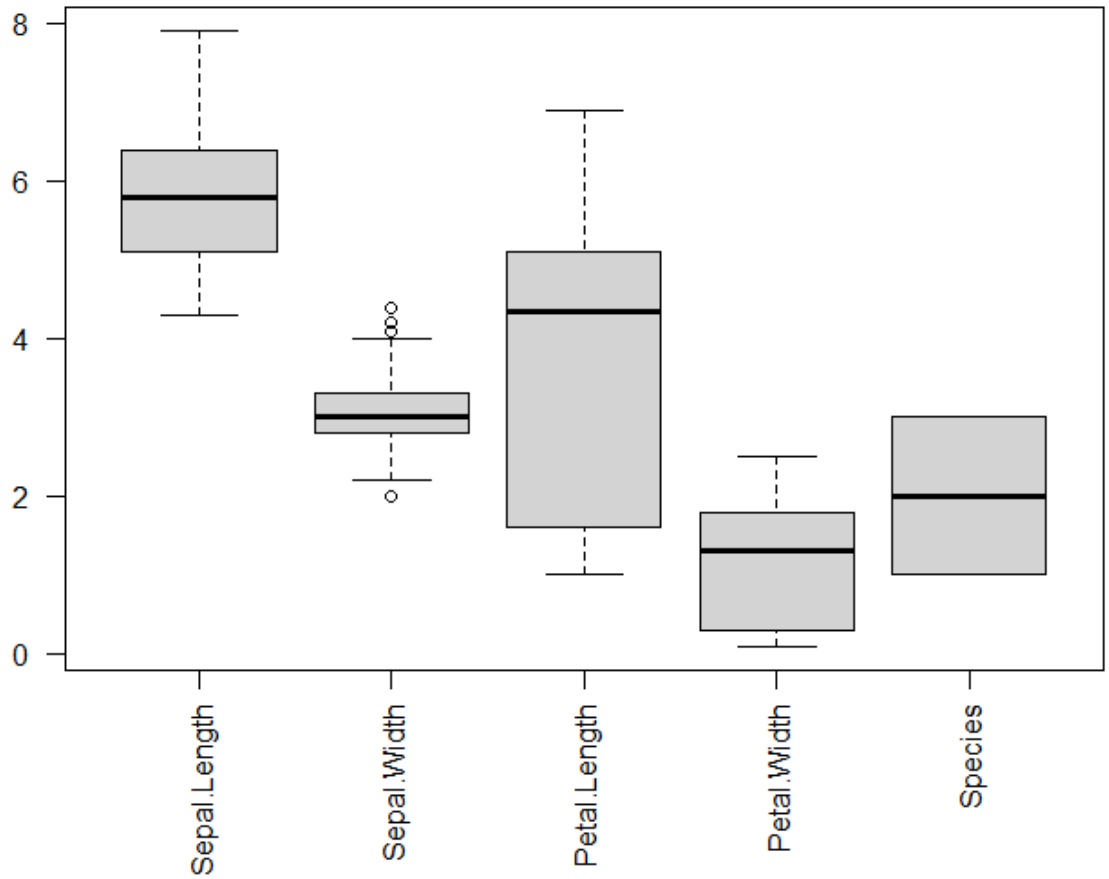

Figure 1: Boxplot summary

Through figure 1 we can see for example the presence of outliers for the feature "Sepal width", most above 1.5 times the 3rd quantile. We can also remark that most "Petal width" values are below the median, same for "Petal width" -even though it is less pronounced-, while "Sepal length" values are more cantered around the median.

This gives us a rough estimate of the distribution of the values for each attribute. But maybe it could be more informative to see the distribution of the values considering each class, since we have labels for each class.

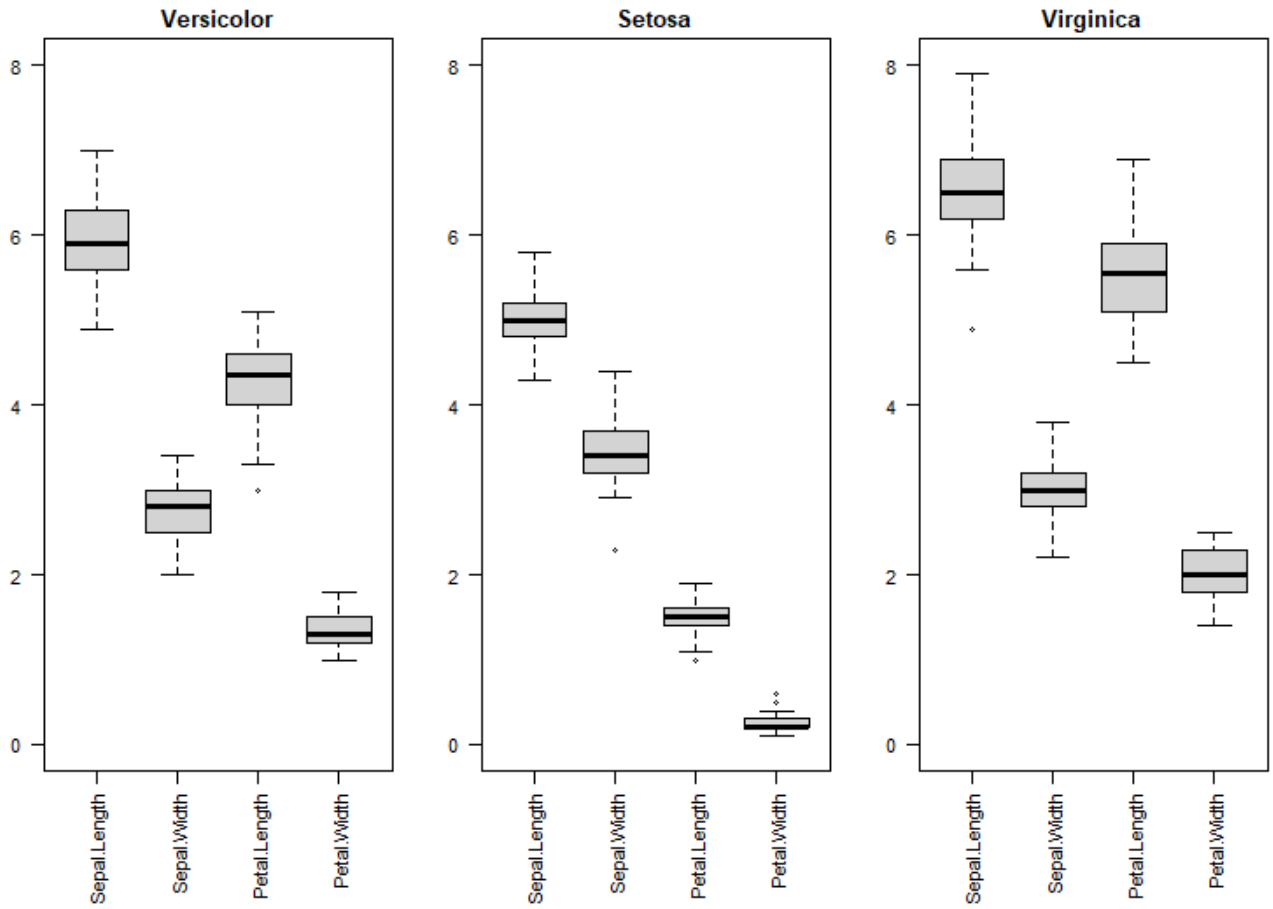

Figure 2: Boxplot by species

When observing the boxplot for each feature and for each class we can notice the subtle differences between the iris species:

*For Sepal length we can see that Virginica has the longest sepals, followed by Versicolor and then Setosa.

*For Sepal width however, it is Setosa that has the biggest width followed by Versicolor and then Virginica.

*For Petal length Virginica has the longest petals followed by Versicolor and then Setosa- that has significantly shorter petals-.

*For Petal width it is again Virginica that comes first, then Versicolor and finally Setosa that is not only the iris that has the smallest petal width but also a small variance for this feature.

After observing this graph, we can estimate that we can use the Petal length feature as one of the features that can help us the most differentiate between the iris species, since the most pronounced difference is seen for this feature.

We can trace the histograms for each class for this feature to observe this more:

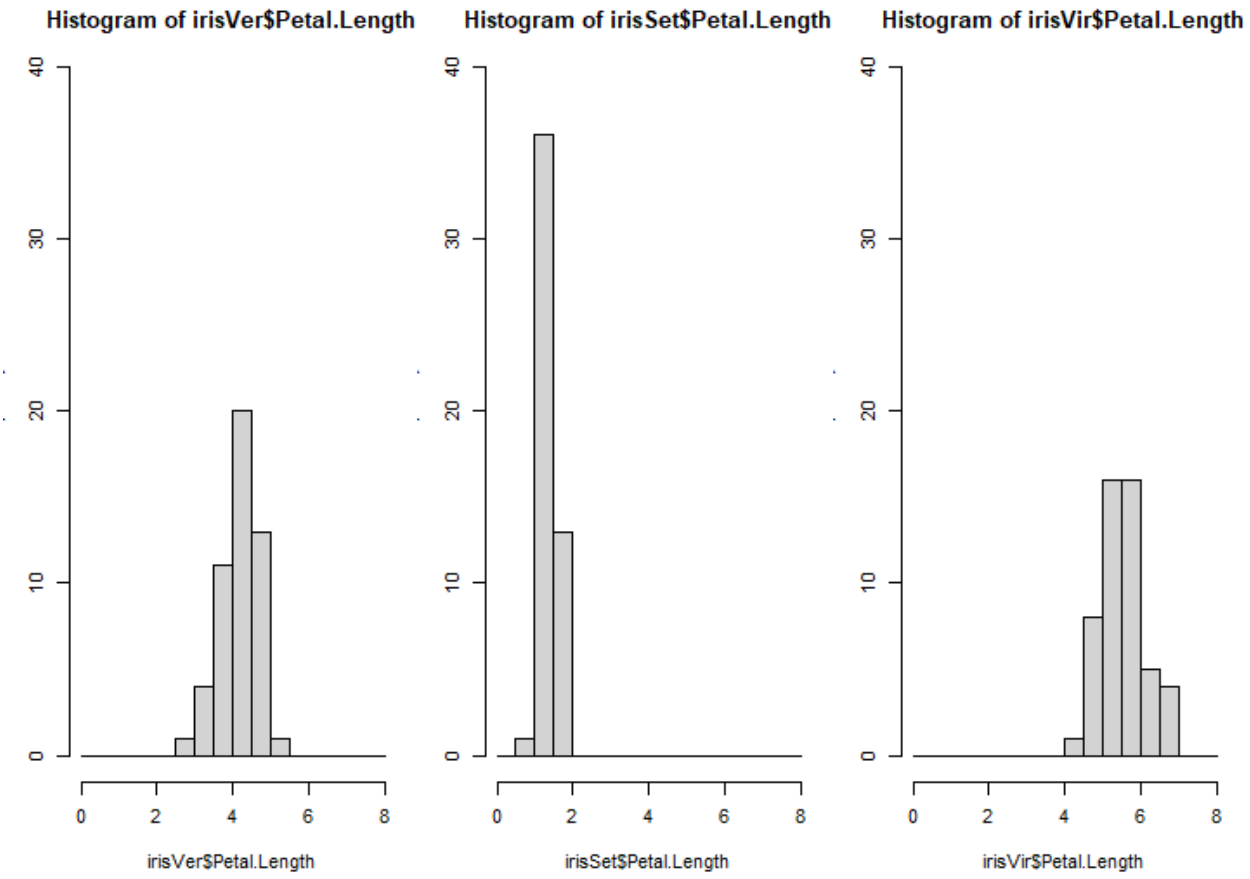

Figure 3: Petal length histograms

As observed earlier, petal length varies from one specie to another: For Versicolor (1st histogram) the petal length is cantered around 4cm, while for Setosa (2nd histogram) it is between 1~2cm and for Virginica it is the longest cantered around 5~6cm. This could be a feature that helps identify iris species.

Correlation between variables:

We then study the correlation between features:

	Sepal. Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000	-0.118	0.872	0.818
Sepal.Width	-0.118	1.000	-0.428	-0.366
Petal.Length	**0.872**	-0.428	1.000	0.963
Petal.Width	**0.818**	-0.366	**0.963**	1.000

Observing the table above we can see that there is positive correlation between Sepal length and Petal length, meaning the longer the sepal is the longer petal is too. The same goes for Sepal length and Petal width as well as an even higher correlation between Petal length and Petal width (0.963).

Scatterplot matrices are very good visualization tools and may help identify correlations or lack of it, the graph bellow accentuates the interpretations given above.

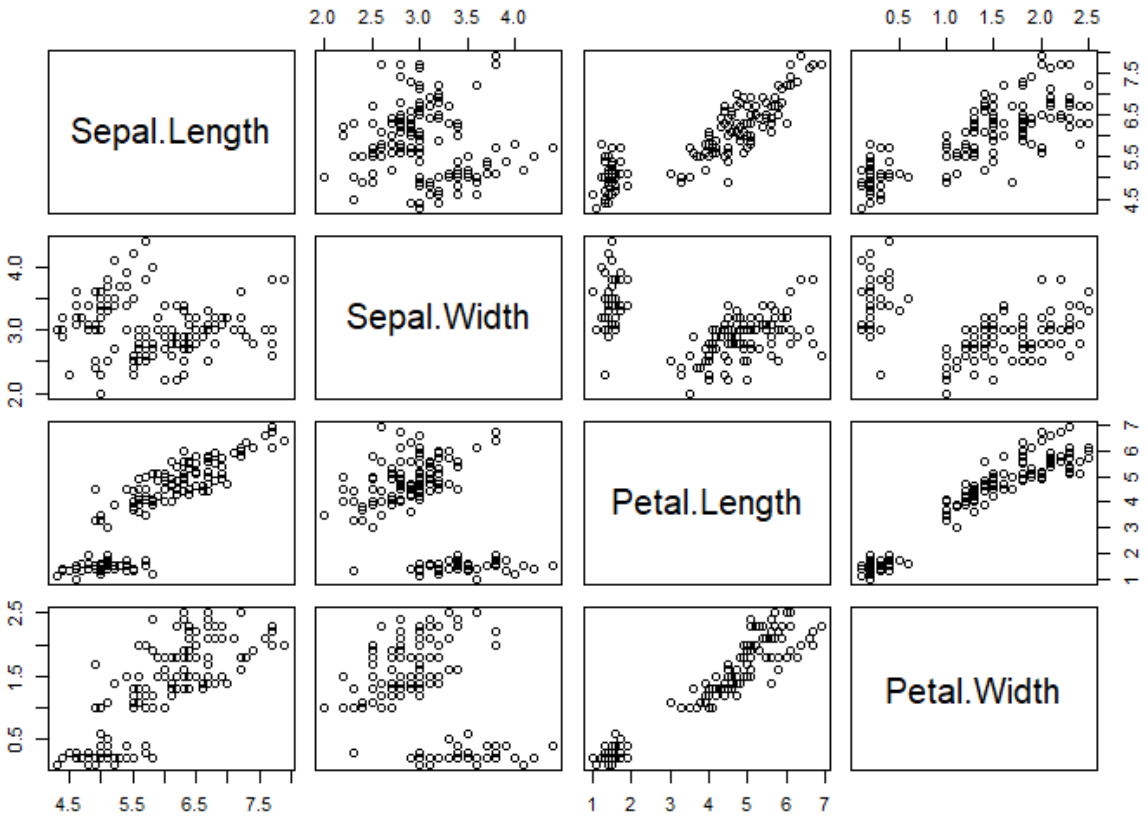

Figure 4: Scatterplot

We can also try to evaluate the correlation depending on the class:

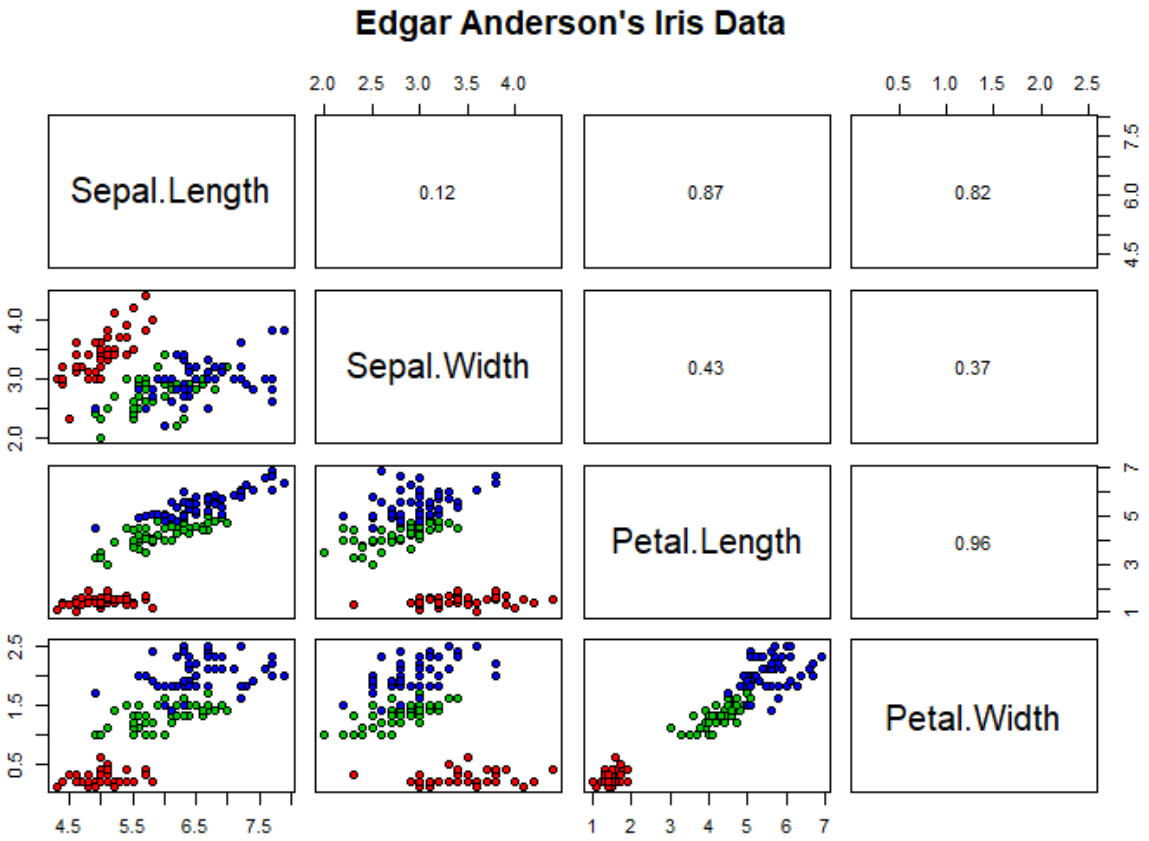

Figure 5: Edgar Anderson plot by class

Figure 5 clearly shows that one particular species, Setosa constitutes the smaller cluster in the low left. The other two species also show a difference in this plot, even though they are not easily separated. This is a very important insight into this dataset.

It looks like most of the variables could be used to predict the species - except that using the sepal length and width alone would make distinguishing Iris versicolor and virginica tricky (green and blue), enforcing the previous result showing that Petal length can be the feature that helps distinguish the two similar species.

Another way to plot a data frame's values to see correlations and values in general are through a parallel coordinate plot.

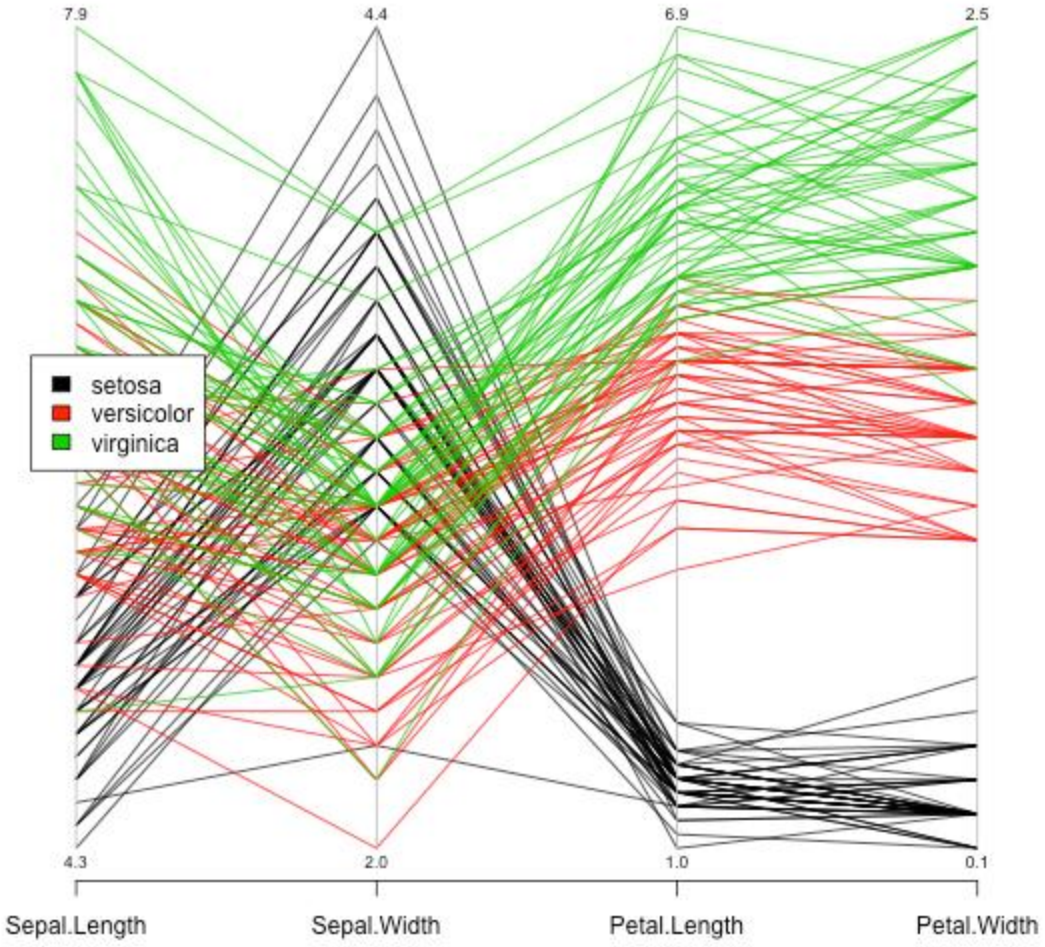

Figure 6: Parallel coordinate plot

Classification with Decision Trees:

Even if we already know the classes for the 150 instances of irises, it could be interesting to create a model that predicts the species from the petal and sepal width and length. One model that is easy to create and understand is a decision tree:

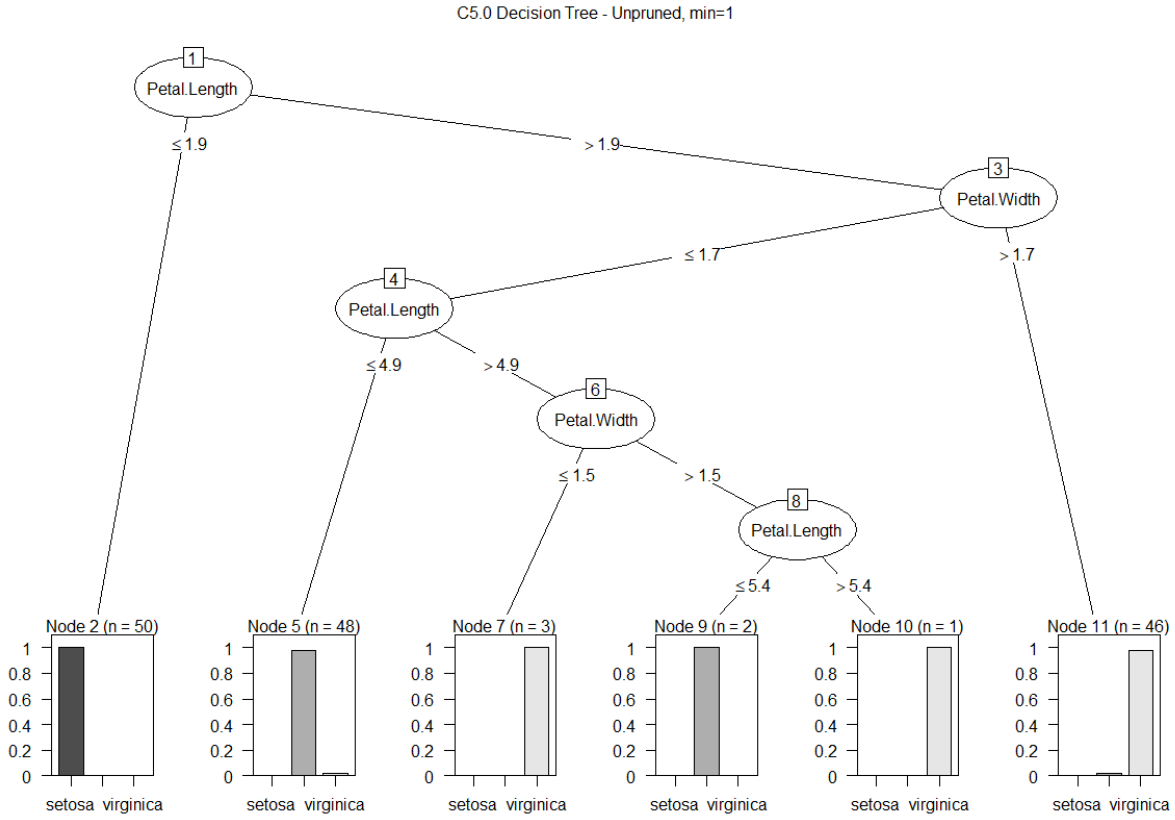

Figure 7: Decision tree

When starting from the top of the decision tree we can observe that Petal length feature is the most decisive feature to classifying iris species, as interpreted earlier. From the bottom we can see the classes: Setosa in dark grey, Versicolor in grey and Virginica in light grey.

Setosa is the one most rightly-classified and just by observing the petal length: in Node 2 we can see all the 50 Setosa irises stemming directly from node 1 for petal length. Versicolor and Virginica on the other hand -having similarities- are slightly harder to classify. Versicolor are divided between Node 5 and Node 9 and again in Node 4 for petal length this feature is again deterministic as through Node 4 (petal length) we find 48 out of the 50 Versicolor. Finally Virginica is divided between Node 11 and Node 7.

We can play with the parameters of the classifier to see better/simpler/more complete/more complex trees. Here's a simpler one:

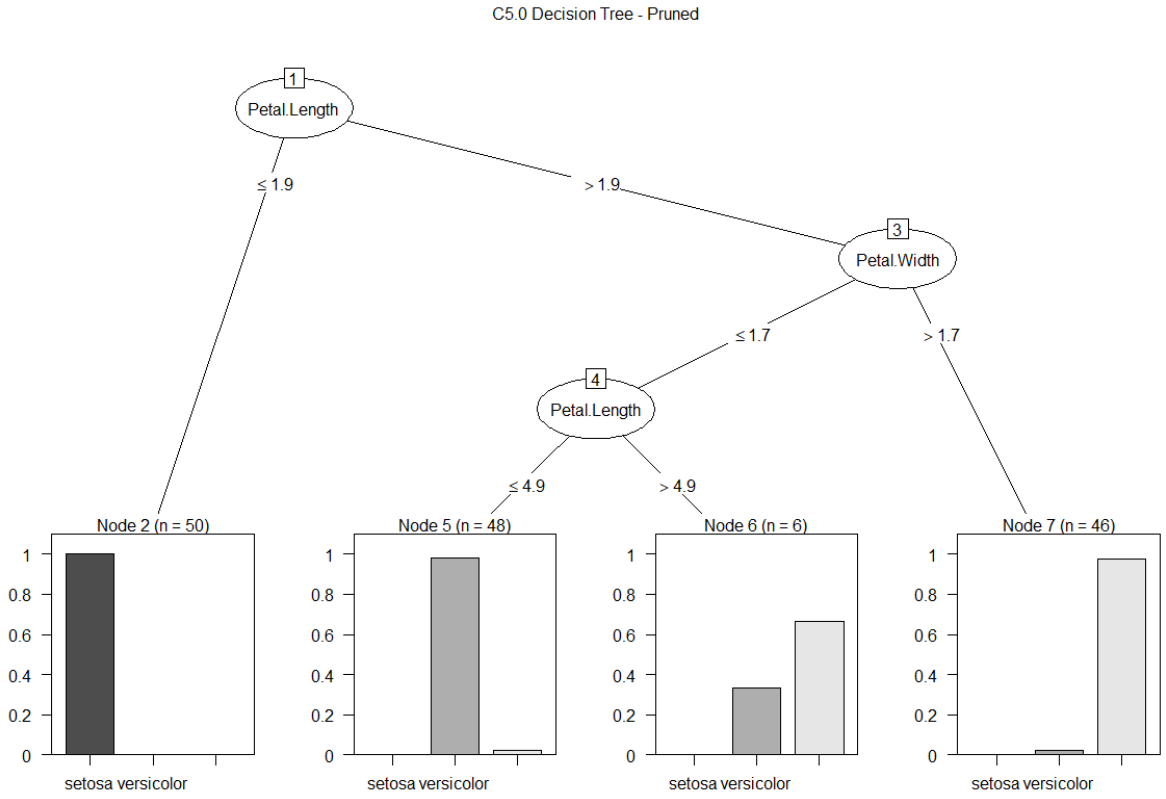

Figure 8: Decision tree simplified

We can see above a simpler decision tree with less nodes and where the misclassified Versicolor and Virginica (6 in total) are all put in one class (Node6).