Can machine Learning be used to identify species of Sorbus

PetraGuy, Imperial College London

February 20 2018

```
#This creates pdf from command line, note, sensitive to ' or "

#Rscript -e "library(knitr); knit('MiniProj2.Rmd')"

#Rscript -e "library(rmarkdown); render('MiniProj2.md')"
```

0.Abstract.

Machine learning was used to separate six species of Sorbus within the subgenus Soraria based on morphological measurements of fruit and leaves. Box plots show that there is considerable overlap of characteristics between the species, but that some species were sufficiently differentiated in one or two characteristics. This was reflected in the modelling where unsupervised clustering algorithms gave inaccurate results but decision tree methods were successful.

1. Introduction - The genus Sorbus.

Sorbus is a member of the Rosaceae family, perhaps the best known species being Sorbus aucuparia, the Rowan or Mountain Ash.

However, there are over 50 species of Sorbus in the UK, 38 of these are vulnerable or critically endangered and most are endemic or native. There are four diploid species, but, as with many Rosaceae, Sorbus produce new apomictic polyploid species. These can also produce viable pollen and can therefore back-cross with other diploid species, (**???**). This results in the large number of genetically unique, stable, clonal communities, which can look very similar to each other. This presents a problem with recording and many Sorbus require expert knowledge to correctly identify to species level because much of the identification depends on comparative knowledge. This tends to dissuade recorders, or encourages records at aggregate level. This is a problem for such an important genus with many endangered plants that could benefit from identification.

Sorbus are grouped into six subgenera, each of which are reasonably easy to identify by recorders with some knowledge, more difficulty arises when identifying plants within these subgenera, and this is where this work has concentrated. In this modelling only the subgenus Soraria has been trialed. This subgenus consists of eight species all similar in appearance to Sorbus intermedia, although only seven species are considered based on the availability of data. These plants are distinguished from other subgenera by having leaves with

rounded lobes which are tomentose beneath and the fruits having fewer lenticels. Perhaps the most noticeable difference between plants within the subgenus, are the larger fruits on S intermedia, the smaller leaves of S minima and the small fruits of S mougeotii.

2 Data and data preparation

The data was provided by Dr T Rich, the Botanical Society of Britain and Ireland expert on Sorbus, and consists of leaf and fruit measurements. For the leaves, the length, width, widest point on the leaf, base angle, number of veins, depth of the lobes, and vein angle have been recorded. For the fruit, the length and the width are used. Due to the variability in leaf size across one plant, the measurements were all carried out in a specific manner described by Rich et al, (Rich 2010). Essentially, repeated measurements of leaves on sterile spurs on the sunlight side of the tree are recorded and averaged over at least ten leaves.

The nature of collection means that the data was sparse. Every plant of every species did not have have complete set of measurements or the same number of measurements. For example, S intermedia had 126 observations but S leyana only had 39. This is due to the rarity S. leyana. S. intermedia is a common plant found throughout the UK in easily accessible places, whilst S leyana is only found in two sites in South Wales, sometimes on the sides of cliffs. In addition, measurements cannot all be collected at the same time. Leaves must be measured when mature, around flowering time, and therefore cannot be measured in conjunction with fruit. Separate trips to re-measure fruit on the same trees may not be possible. This has lead to a sparse data set in which not all morphological characteristics were available for every plant. S intermedia records are an example. Of 122 records, 72 are purely for fruit measurements and the remaining 50 purely for leaf measurements, and these occur on different plants If imputation was carried out, 59% of the leaf measurements would be imputed. This would reduce the effectiveness of some algorithms. For example, in kNN, if you increase the frequency of the neighbours in the S. intermedia group, it is more likely that a member of a different group will be close to that neighbour. Therefore, the sparsity was handled by reallocating measurements. For example, the 50 leaf measurements for S. intermedia were assigned to 50 fruit measurements and the excess 22 were not used. In some cases, where there were only a few additional rows of incomplete data, median imputation was carried out.

Although it seems dubious to assign records from one plant to another, in this analysis this was felt to be acceptable for two reasons. Firstly, this an exploration of a new technique for biological recording. It is not currently being proposed as a complete and accurate method for species identification at this stage. Secondly, the clonal nature of these plants implies that we would expect a great deal of similarity within a species. The variation within the species is more likely to come from the variety of leaf sizes which can be found on one plant, and these are controlled for, although they cannot be eliminated, when the data is collected. However, if these plants are phenotypically very plastic, these assumptions may be invalid. The range of leaf sizes within each plant was not available, but a comparison of the variation within each plant and between all the plants of each species would be useful here. That analysis would demonstrate whether each record needs to be complete or not.

This procedure also has the benefit of producing a data set with no missing values, and some of the machine learning algorithms used here had no method for dealing with these, hence they must be removed before modelling. Some machine learning algorithms are sensitive to scale in the data, for example, k nearest neighbours and k-means, therefore the data was also standardized and both standardised and non-standardised data was modelled for kmeans and knn. The non-standardised data only was used in he regression tree model because these do not require standardised data.

Some machine learning algorithms require train and test sets. The model is fitted to a subset of the data – the training set, and its performance evaluated using new data – the test set. This is to avoid over fitting and to improve the predictive power of the models. The data can be split by selecting a random sample of, for example, 70% of the data. However, this might be problematic with this data because it is unbalanced. Therefore, a random stratified sampling system was carried out. Each species, which has a different number of entries, was split into 70/30 train test sets. This should reduce the bias of the model whilst not over fitting.

In addition, cross fold validation was also used to increase accuracy. This technique repeatedly creates train test sets as described above and averages the model performance metrics across all the folds. This gives a more robust estimate of the model accuracy because it is less dependent on the choice of the data for the train and test sets.

3. Modelling

3.1Model performance metrics

In classification models, the correct and incorrect values assigned to each class are known, and these can be used to evaluate the model.

Accuracy is the number of correct values divided by total number of items evaluated.

Precision is the ratio of true positives to false positives in each species, so there will be precision for each species. Precision tells you how accurately the algorithm is correctly placing species, a low precision tells you that many other species are lumped with the correct species.

Sensitivity is the true positive rate of a class. Sensitivity tells you how good the classes are, low number tells you the correct species have been put in other, incorrect, classes.

For clustering models, well defined clusters represent better models and the ratio of within cluster sum of squares to total sum of squares was used. For well defined, compact clusters the ratio will be small.

For all the models, because we do have information about the clusters, that is, we know the species, we can use accuracy, precision and sensitivity and therefore we can compare the three modelling methods using the same metrics.

3.2 Modelling methods.

Three machine learning methods were used k-means, hierarchical clustering and a decision tree. The first two being unsupervised clustering techniques and the third a supervised classification algorithm.

3.2.1Decision tree.

Variables are used to make binary decisions as whether data points are part of a group or not. Decisions are made based on whether the information after the decision, i.e., the separation of the groups, is increased or decreased. The final classes would ideally contain only the items of a single species, this will not be the case due to noise within the data. The model here is be represented by the logical processes followed to reach the final classes. The rpart package was used for the decision tree (**???**). The rpart package also produces a plot for the regression tree which summarises the choices at each node and provides a useful tool for identifying and assessing the probability of correct identification of a species.

3.2.2.K-means

Kmeans, which is part of base R, is an unsupervised clustering technique. Even though we do know the identity of the instances in the data, this is not used in the model. Instead, the data is grouped into clusters where the aim is to make the items within each cluster similar, whilst each cluster is as dissimilar as possible from other clusters. This is similar to a classification technique except the classes to which the items belong is to specified. In clustering, no information is needed about the objects and there is no right or wrong, so in that sense, this problem is not strictly a clustering problem. We know what species a sample belongs to and we do not want it allocated to another cluster. However, it is a useful technique for examining the data and

revealing patterns within the data. The k in k means refers to the number of clusters to be used, which we specified as seven – the number of species.

In k-means k centroids are randomly assigned to the data. The data points are then assigned to the closest centroid, resulting in k clusters. The centroid is then moved to the average location of the data-points in its cluster. This process is repeated until the centroid position is stable, or the maximum number of iterations has occurred. If repeating the k-means function results in different clusters, which can be seen in differences in accuracy, precision and the numbers of true positives, it can be assumed that the algorithm is not efficient at separating clusters. Since the number of clusters is known, repeating the algorithm and examining the true positives will indicate the success of the model.

In order to explore different k-means models the algorithm was also repeated on imputed, standardised and semi-standardised data. In semi-standardised data only the variables which were orders of magnitude larger were manipulated. The standardize function in R was used to subtract the means and divide by the standard deviation. The MacQueen method gave the highest accuracy and was used for all the calculations.

3.2.3. Hierarchical Clustering.

Instead of randomly assigning k centroids, hierarchical clustering assigns each data-point to a single cluster. The distance between the clusters is calculated and the closest two points are aggregated into a new cluster, so the clusters decrease by one. The process is repeated until all items are clustered into one cluster. The clusters can be cut at k and the members can be examined. Hierarchical clustering was explored using different distance calculation methods.

Hierarchical clustering is again unsupervised, but since we know the members of each cluster, we can compare the clusters to the original data and calculate accuracy, precision and sensitivity. The hclust function is part of the stats package which is usually included in base R.

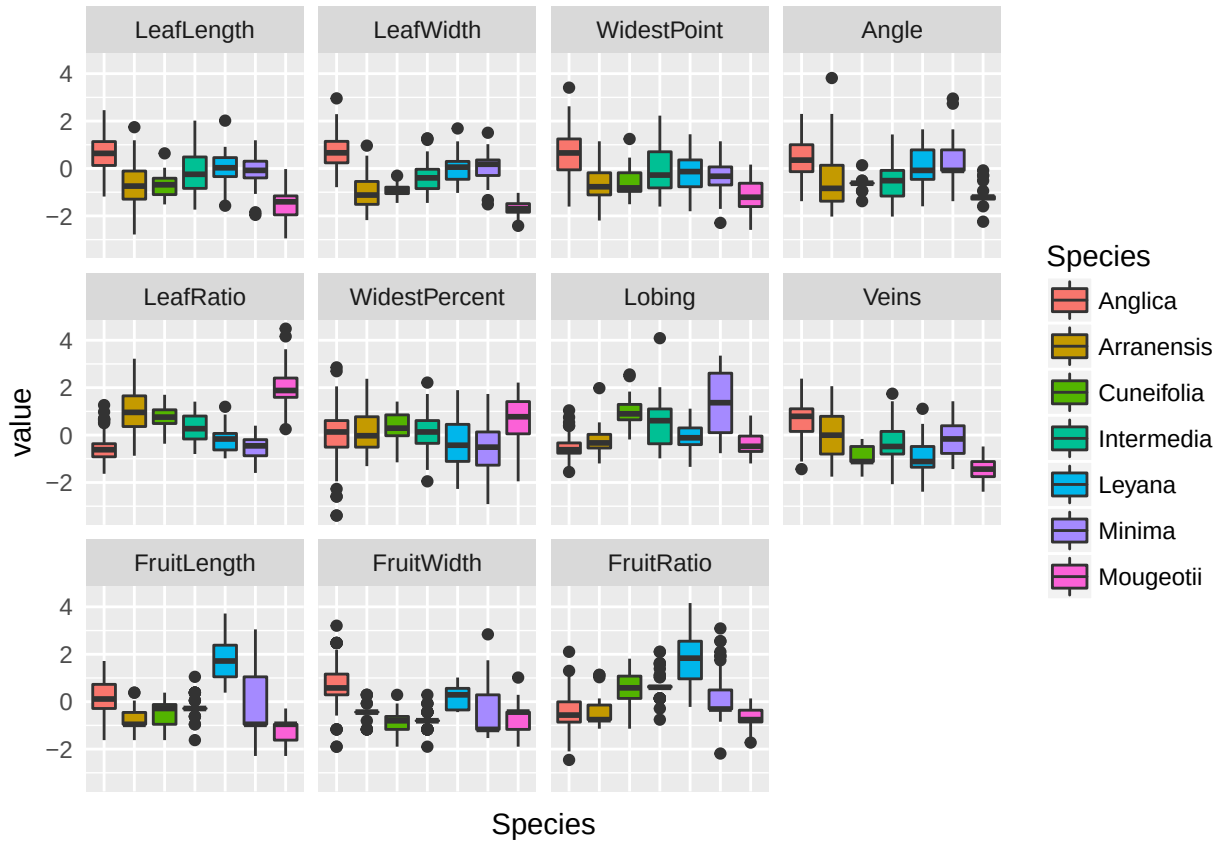
3.4 Computing languages

R was the main language used in this project, although there is no reason, in terms of functionality, why Python could not be used, especially at the more simplistic level of modelling carried out here. A large benefit in R was that it can easily be used in conjunction with R markdown which then provide a mechanism for easily producing pdf documents with an interactive document. In addition, the data was provided by, and the results prepared for, members of the ecological community, where R is the most common package being used. Python was used for some data preparation in order to full fill the criteria of the project, but R would have been equally suitable. R markdown was used in preference to Latex since it provides the same functionally as Latex but with the added benefit of being a dynamic document that is commonly used by other researchers in ecology.

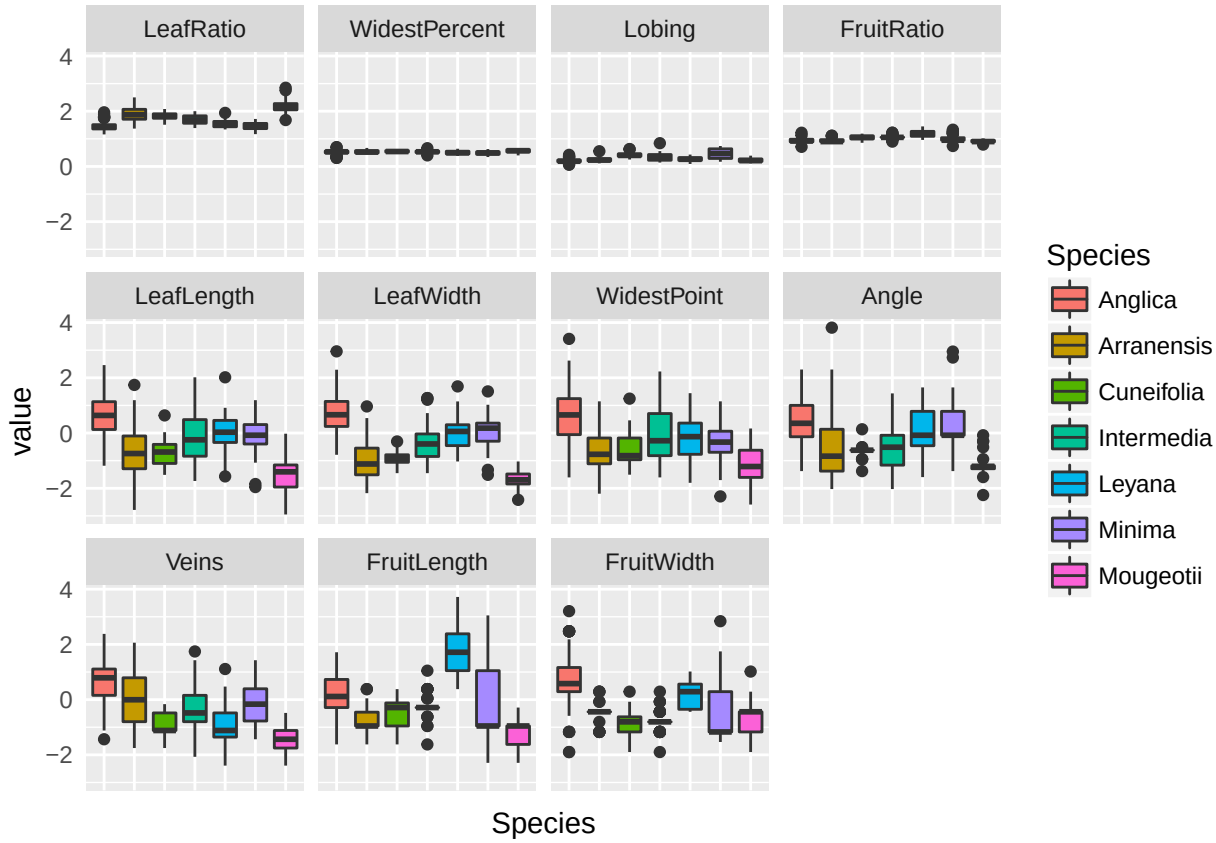
4.Data exploration.

In order for machine learning algorithms to work accurately the groups should be separated into clearly defined clumps with very little overlap. Box plots show the similarity between the variables and how much the variables overlap.

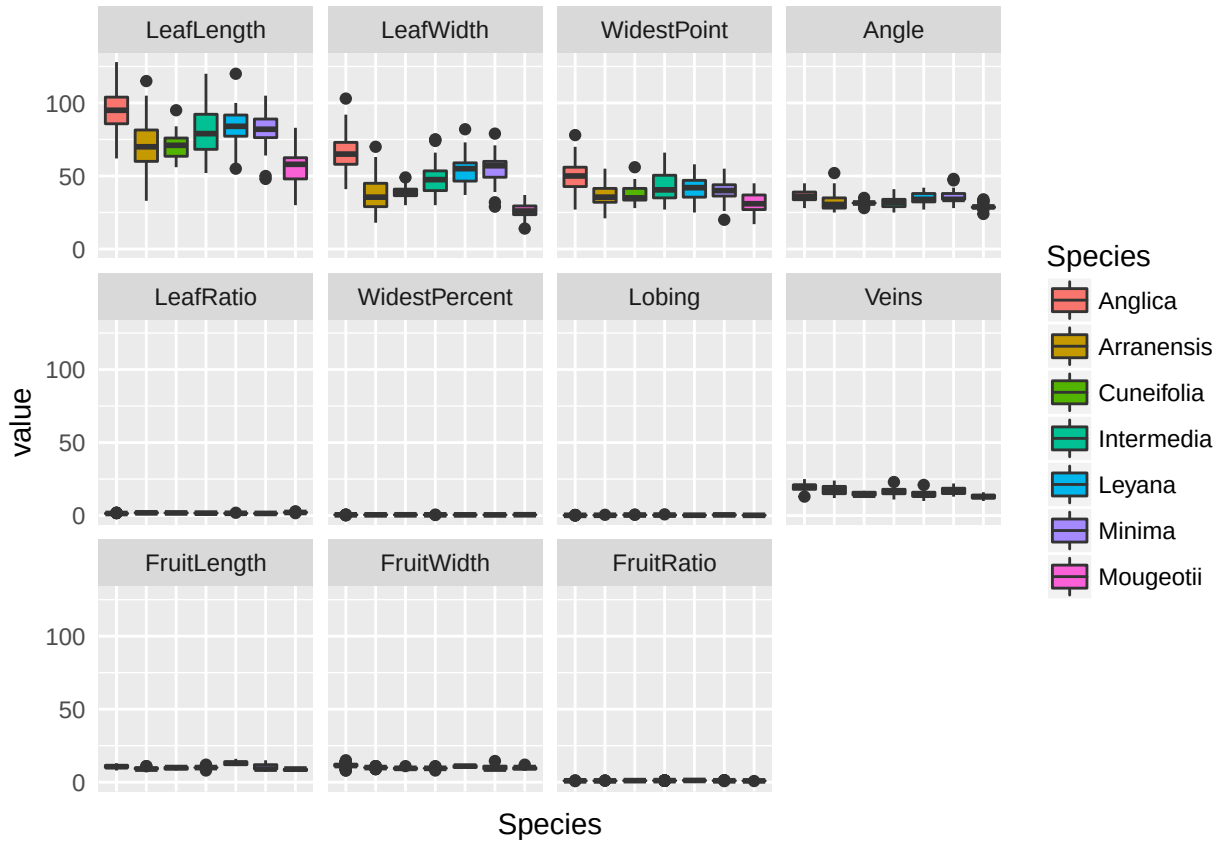
The box plots are presented for the imputed semi-standardised and fully standardised data and show how the scale and overlap could be an issue for the clustering algorithms. The plots also show that despite the overlap, certain features clearly differentiate certain species. For instance, fruit width would separate S. anglica, and then fruit length would subsequently separate S leyana. This suggests that a decision tree algorithm could be successful. The plots also show that more data preparation might need to be employed, for example, scaling some of the variables instead of standardising.


LeafLength
LeafWidth
WidestPoint
Angle
LeafRatio
WidestPercent
Lobing
Veins
FruitLength
FruitWidth
FruitRatio
value
4
2
0
−2
Species
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii

156


LeafRatio
WidestPercent
Lobing
FruitRatio
LeafLength
LeafWidth
WidestPoint
Angle
Veins
FruitLength
FruitWidth
value
4
2
0
−2
Species
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii

157



5.

Results 5.1 Kmeans

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
unscaled	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
semi-scaled	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
fullyscaled	0.42	0.42	0.43	0.42	0.42	0.42	0.42	0.42	0.42	0.43

The table shows the within cluster sum of squares to between cluster sum of squares across the 10 repeats. The ratio is better for the unscaled data and a ratio of 0.2 is often considered as acceptable.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
unscaled	0.19	0.17	0.11	0.05	0.07	0.14	0.22	0.07	0.20	0.26
semi-scaled	0.05	0.08	0.09	0.23	0.13	0.11	0.18	0.29	0.11	0.04
fullyscaled	0.21	0.14	0.04	0.06	0.29	0.20	0.13	0.07	0.07	0.08

The precision is different on each run which implies that the algorithm is not successfully grouping the data into the same clusters.

The next tables show the percentage of each species correctly allocated to its cluster on each of the ten repeats for the For example, the top row from left to right, gives the true positive rate for S. anglia on each subsequent run on the kmeans algorithm.

unscaled

	1	2	3	4	5	6	7	8	9	10
Anglica	21.25	20.62	11.88	3.12	0.00	12.50	26.25	0.00	21.25	31.87
Cuneifolia	40.00	0.00	0.00	22.00	2.00	4.00	30.00	6.00	20.00	20.00
Intermedia	0.00	52.63	0.00	0.00	0.00	26.32	52.63	0.00	0.00	5.26
Leyana	2.08	6.25	10.42	2.08	31.25	4.17	6.25	31.25	6.25	2.08
Minima	3.33	3.33	6.67	3.33	36.67	36.67	36.67	23.33	3.33	3.33
Mougeotii	32.00	32.00	8.00	4.00	0.00	4.00	4.00	4.00	46.00	46.00

```
Arranensis  0.00  0.00 52.17  0.00  0.00 43.48  0.00  4.35 21.74 43.48
```

semi-scaled

```
                1     2     3     4     5     6     7     8     9    10
Anglica      1.25  1.25  5.62 21.88  1.25  1.25 20.62 30.63  6.25  6.25
Cuneifolia   0.00  8.00 22.00 22.00  4.00 30.00  0.00 38.00 44.00  0.00
Intermedia   0.00  0.00 47.37 52.63  0.00  0.00 47.37  0.00  0.00  0.00
Leyana       0.00 45.83  0.00 18.75 45.83  0.00 45.83 16.67 10.42  0.00
Minima       3.33  0.00 10.00 66.67  3.33  0.00  3.33 83.33 10.00 13.33
Mougeotii   32.00  2.00  4.00  2.00 48.00 48.00  4.00 16.00  2.00  2.00
Arranensis   4.35  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
```

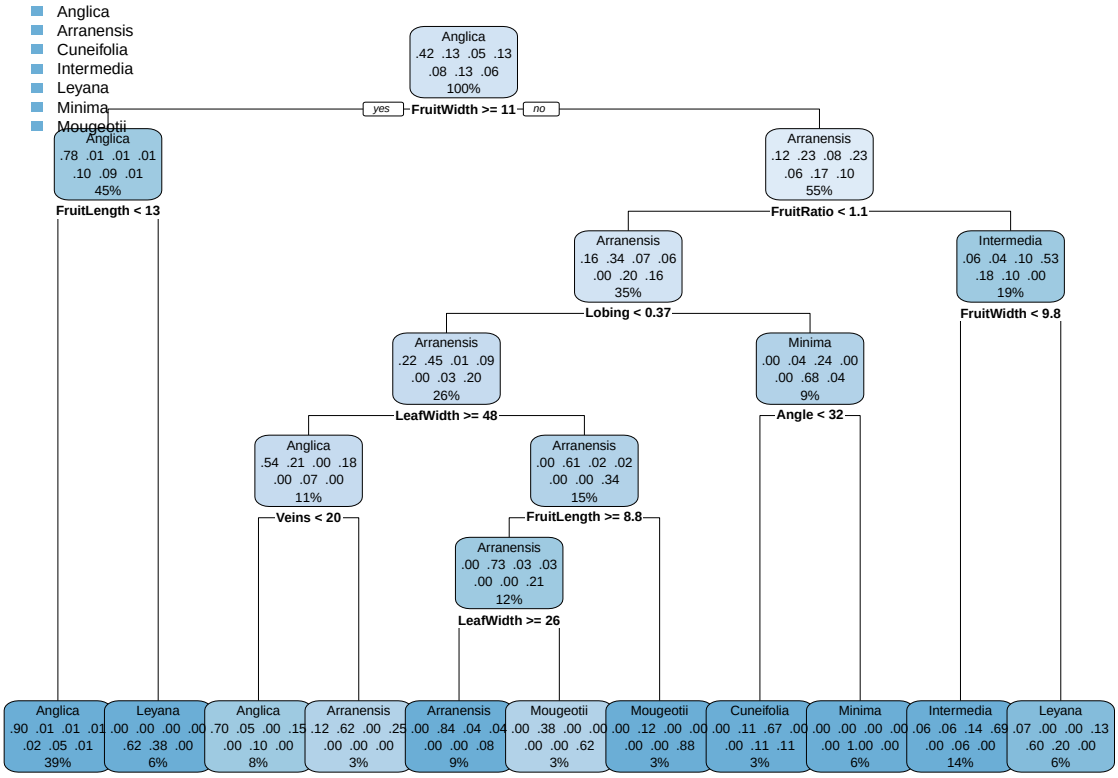
scaled

```
                 1     2     3     4     5     6      7     8  9    10
Anglica      31.87 19.38  1.88  0.00 48.12 26.25  16.88  0.00  0  0.00
Cuneifolia    0.00  0.00  2.00  4.00  0.00  0.00   6.00  0.00 44  4.00
Intermedia    0.00 94.74  0.00  0.00  0.00  0.00 100.00  0.00  0  5.26
Leyana        8.33  4.17 22.92  2.08 62.50 68.75   4.17 58.33  0 52.08
Minima        0.00  6.67  6.67  6.67  6.67  0.00   0.00  0.00  0  0.00
Mougeotii     0.00  0.00  0.00 32.00  6.00  4.00   0.00  0.00  6  6.00
Arranensis  100.00  0.00  0.00  0.00  0.00  0.00   0.00  0.00  0  0.00
```

Accuracy obtained using the different distance calculations

```
Distance Method "euclidean" "maximum" "manhattan" "canberra" "minkowski"
Accuracy        "0.3"       "0.18"    "0.29"      "0.42"     "0.3"
```

```
             cluster
                1   2   3   4   5   6   7
  Anglica     108  41  11   0   0   0   0
  Arranensis    7  14   0   0  22   0   7
  Cuneifolia    0   1   0   8  10   0   0
  Intermedia   13   2   2  22   8   1   0
  Leyana        4  17   6   0   0   3   0
  Minima        1   8   9  29   0   3   0
  Mougeotii     0   0   0   0  12   0  11
```



The decision tree produces some very accurate nodes. 100 % for S. minima, 90% for S. anglica,

The accuracy for the decision tree is

0.68

The sensitivies for the decision tree model are shown in the table below

```
     Species      Sensitivity
[1,] "Anglica"    "0.85"
[2,] "Arranensis" "0.47"
[3,] "Cuneifolia" "0.17"
[4,] "Intermedia" "0.71"
[5,] "Leyana"     "0.67"
[6,] "Minima"     "0.6"
[7,] "Mougeotii"  "0.43"
```

The precisions for the decision tree model are shown in the table below

```
     Class              Precision
[1,] "class_Anglica"    "0.79"
[2,] "class_Arranensis" "0.5"
[3,] "class_Cuneifolia" "0.5"
[4,] "class_Intermedia" "0.67"
[5,] "class_Leyana"     "0.4"
[6,] "class_Minima"     "1"
[7,] "class_Mougeotii"  "0.43"
```

The confusion matrix details exactly how the species were placed.

The final column is the sum of species in the test set.

```
           Anglica Arranensis Cuneifolia Intermedia Leyana Minima
Anglica         41          3          0          0      3      0
```

```
Arranensis       4          7          0          0      1      0
Cuneifolia       0          1          1          4      0      0
Intermedia       2          0          1         10      1      0
Leyana           3          0          0          0      6      0
Minima           1          0          0          1      4      9
Mougeotii        1          3          0          0      0      0
           Mougeotii
Anglica             1 48
Arranensis          3 15
Cuneifolia          0  6
Intermedia          0 14
Leyana              0  9
Minima              0 15
Mougeotii           3  7
```

Rich, Houston, T. 2010. *Whitebeams, Rowans and Service Trees of Britain and Ireland.* BSBI.