Can machine Learning be used to identify species of Sorbus

PetraGuy, Imperial College London

February 20 2018

Pl. 113. Sorbier des Oiseleurs. Sorbus aucuparia L.

0. Abstract.

Machine learning was used to separate seven 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 backcross with other diploid or polyploid 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 trialled. 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 lenticles. 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 [ ]. 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 dataset 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 to aid biological recording; it is not being proposed as a complete and accurate method for species identification. 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 dataset 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 models carried out on standardized and non-standardized 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. For these types of models accuracy, precision and sensitivity are used to evaluate the model.

Accuracy is 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 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 data, despite using unsupervised learning methods, we do know the identification of the species, this means that we can calculate accuracy, precision and sensitivity for all the models and compare them using the same metric.

Confusion matrices, which summarise the the frequencies of the species allocated to different classes and clusters can also be produced.

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.

3.2.1.Decision 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. []

Accuracy, precision, sensitivity and a confusion matrix were used to evaluate the decision tree, as well as a decision tree plot which summarises the binary choices used at each node. since standardization should not effect a decision tree, only the unstandardized data was modelled.

3.2.2. K-means

Kmeans 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 non-standarized data. 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.

Accuracy, precision and sensitity were used to evaluate the algorithm, as well as between cluster to within cluster ratio. The consistency of the algorithm over ten was was examined and the percentage of true positives on each run calcualted. Since there were repeated runs, displaying a confusion matrix in each case would be visually difficult to examine, so they were not used in this case.

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.

Accuracy, precision, sensitivity and a confusion matrix were used to evaluate the hierarchical clustering technique.

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 as it provides the same functionality as Latex, allowing the use of latex commands directly within the document, 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.

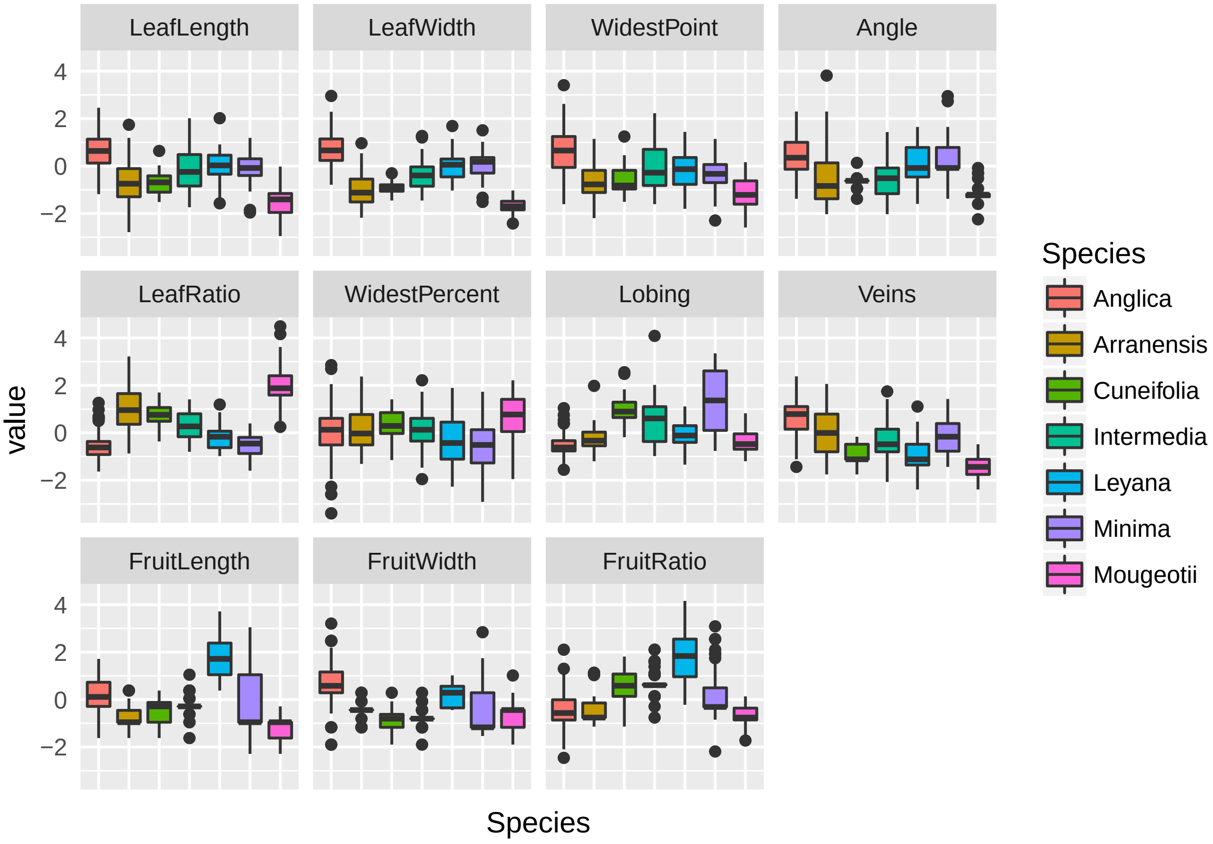

Figure 1: Box plots for standardized data

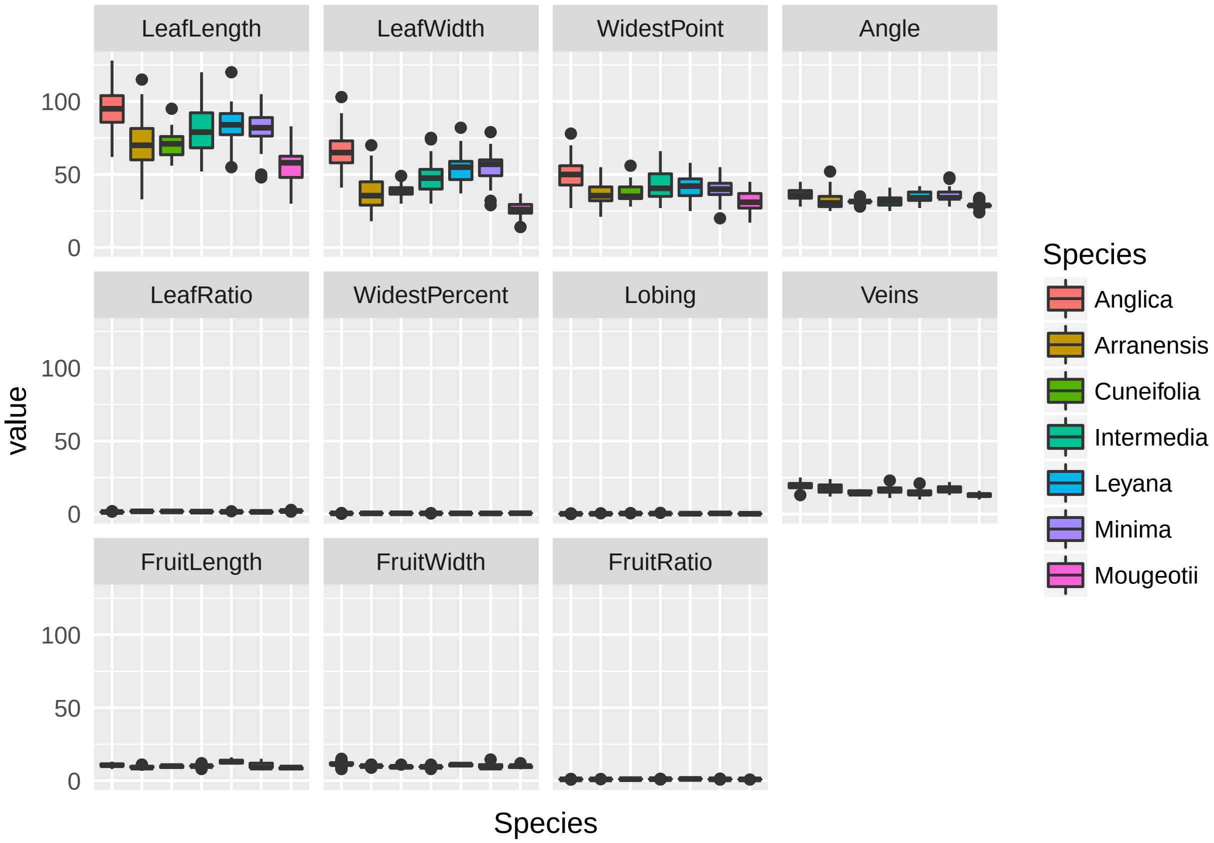

Figure 2: Box plots for un-standardized data

5. Results

5.1 Kmeans

Table 1: Within cluster to between cluster ratio

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
unstandardized	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
standardized	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42

Table 1 shows the within cluster sum of squares to between cluster sum of squares across the 10 repeats. The ratio is better for the unstandardized data.

Table 2: Accuracy

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
unstandardized	0.22	0.25	0.10	0.08	0.31	0.07	0.12	0.03	0.07	0.12
standardized	0.13	0.23	0.33	0.09	0.35	0.10	0.02	0.17	0.08	0.31

The accuracy shown in table 2 is different on each run which implies that the algorithm is not successfully grouping the data into the same clusters. The mean accuracy is higher for the standardized data, but its low value, its fluctuation and its range imply that the model is not useful.

The following 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.

Table 3: Percentage of true positives for non-standarized data

	1	2	3	4	5	6	7	8	9	10
Anglica	21.25	24.38	3.12	0.00	31.87	3.12	21.25	0.00	3.12	12.50
Cuneifolia	34.00	40.00	0.00	6.00	40.00	8.00	4.00	0.00	2.00	22.00
Intermedia	0.00	5.26	21.05	21.05	5.26	0.00	21.05	21.05	21.05	0.00
Leyana	29.17	10.42	6.25	31.25	29.17	2.08	6.25	4.17	6.25	20.83
Minima	3.33	33.33	3.33	23.33	23.33	3.33	6.67	10.00	36.67	3.33
Mougeotii	0.00	42.00	48.00	0.00	46.00	34.00	0.00	8.00	4.00	8.00
Arranensis	73.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The results again show that the algorithm is not consistently allocating species to the correct cluster. On some runs, it is very accurate for some species, but not necessarily for all the others.

Table 4: Percentage of true positives for standarized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0.00	51.88	31.87	18.75	53.75	0.00	0.00	24.38	0.00	53.12
Cuneifolia	6.00	0.00	4.00	0.00	30.00	48.00	0.00	32.00	2.00	6.00
Intermedia	94.74	5.26	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leyana	6.25	2.08	4.17	0.00	62.50	18.75	6.25	8.33	2.08	62.50
Minima	83.33	0.00	90.00	0.00	6.67	3.33	0.00	6.67	6.67	0.00
Mougeotii	0.00	6.00	48.00	8.00	0.00	6.00	6.00	8.00	50.00	0.00
Arranensis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Then on other runs its is completely inaccurate. The standardized data shows better results than the non standardized data.

In order to compare this model to hierarchical clustering and decision tree, the precision and sensitivity were also calculated.

Table 5: Precision of kmeans with non standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0.21	0.24	0.03	0	0.32	0.03	0.21	0	0.03	0.12
Cuneifolia	0.34	0.4	0	0.06	0.4	0.08	0.04	0	0.02	0.22
Intermedia	0	0.05	0.21	0.21	0.05	0	0.21	0.21	0.21	0
Leyana	0.29	0.1	0.06	0.31	0.29	0.02	0.06	0.04	0.06	0.21
Minima	0.03	0.33	0.03	0.23	0.23	0.03	0.07	0.1	0.37	0.03
Mougeotii	0	0.42	0.48	0	0.46	0.34	0	0.08	0.04	0.08
Arranensis	0.74	0	0	0	0	0	0	0	0	0

Table 6: Precision of kmeans with standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0	0.52	0.32	0.19	0.54	0	0	0.24	0	0.53
Cuneifolia	0.06	0	0.04	0	0.3	0.48	0	0.32	0.02	0.06
Intermedia	0.95	0.05	1	0	0	0	0	0	0	0
Leyana	0.06	0.02	0.04	0	0.62	0.19	0.06	0.08	0.02	0.62
Minima	0.83	0	0.9	0	0.07	0.03	0	0.07	0.07	0
Mougeotii	0	0.06	0.48	0.08	0	0.06	0.06	0.08	0.5	0
Arranensis	0	0	0	0	0	0	0	0	0	0

The precision and sensitivity also demonstrate the instability of the model. Although the standardized data in both cases sometimes gives very good results, this is not consistent or repeatable.

Table 7: Sensitivity of kmeans with non standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0.21	0.24	0.03	0	0.32	0.03	0.21	0	0.03	0.12
Cuneifolia	0.34	0.4	0	0.06	0.4	0.08	0.04	0	0.02	0.22
Intermedia	0	0.05	0.21	0.21	0.05	0	0.21	0.21	0.21	0
Leyana	0.29	0.1	0.06	0.31	0.29	0.02	0.06	0.04	0.06	0.21
Minima	0.03	0.33	0.03	0.23	0.23	0.03	0.07	0.1	0.37	0.03
Mougeotii	0	0.42	0.48	0	0.46	0.34	0	0.08	0.04	0.08
Arranensis	0.74	0	0	0	0	0	0	0	0	0

Table 8: Sensitivity of kmeans with standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0	0.52	0.32	0.19	0.54	0	0	0.24	0	0.53
Cuneifolia	0.06	0	0.04	0	0.3	0.48	0	0.32	0.02	0.06
Intermedia	0.95	0.05	1	0	0	0	0	0	0	0
Leyana	0.06	0.02	0.04	0	0.62	0.19	0.06	0.08	0.02	0.62
Minima	0.83	0	0.9	0	0.07	0.03	0	0.07	0.07	0
Mougeotii	0	0.06	0.48	0.08	0	0.06	0.06	0.08	0.5	0
Arranensis	0	0	0	0	0	0	0	0	0	0

5.2 Hierarchical Clustering

Table 9: Accuracy obtained in hierarchical clustering using different distance metrics for non standarized data

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

Table 10: Accuracy obtained in hierarchical clustering using different distance metrics for standardized data

Distance Method	euclidean	maximum	manhattan	canberra	minkowski
Accuracy	0.41	0.14	0.36	0.25	0.41

The table above shows that the Canberra metric gives the most accurate results in the non-standardized data. The Euclidean and Minkowski methods give slightly worse worse results but the best for the standardized data.

The confusion matrix above shows that the method successful clusters some species, such as S.

Table 11: Confusion matrix for Canberra method using un standardized data

	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

anglica, many other species are dispersed across the cluster.

Table 12: Confusion matrix for Canberra method with standardized data

	1	2	3	4	5	6	7
Anglica	52	17	27	6	10	43	5
Arranensis	5	2	2	12	1	1	27
Cuneifolia	0	0	0	1	0	0	18
Intermedia	6	2	0	10	0	7	23
Leyana	1	2	0	14	8	1	4
Minima	0	2	12	12	13	2	9
Mougeotii	0	0	0	3	0	0	20

The confusion matrix for the standardized data shows worse results than for the non-standardized data.

Table 13: Precision for standardized and non-standardized data

	Standardized	Unstandardized
Class1	0.69	0.81
Class2	0.08	0.17
Class3	0	0
Class4	0.04	0.37
Class5	0.05	0
Class6	0.05	0.43
Class7	0.58	0.61

The precision and sensitivity is not consistent between classes or species. The precision is higher in all cases for the non-standardized data, but the sensitivity is sometimes higher in standardized data.

Table 14: Precision for standardized and non-standardized data

	Standardized	Non-standardized
Anglica	0.68	0.32
Cuneifolia	0.28	0.04
Intermedia	0	0
Leyana	0.46	0.21
Minima	0	0.27
Mougeotii	0.06	0.04
Arranensis	0.48	0.87

5.3 Decision Tree

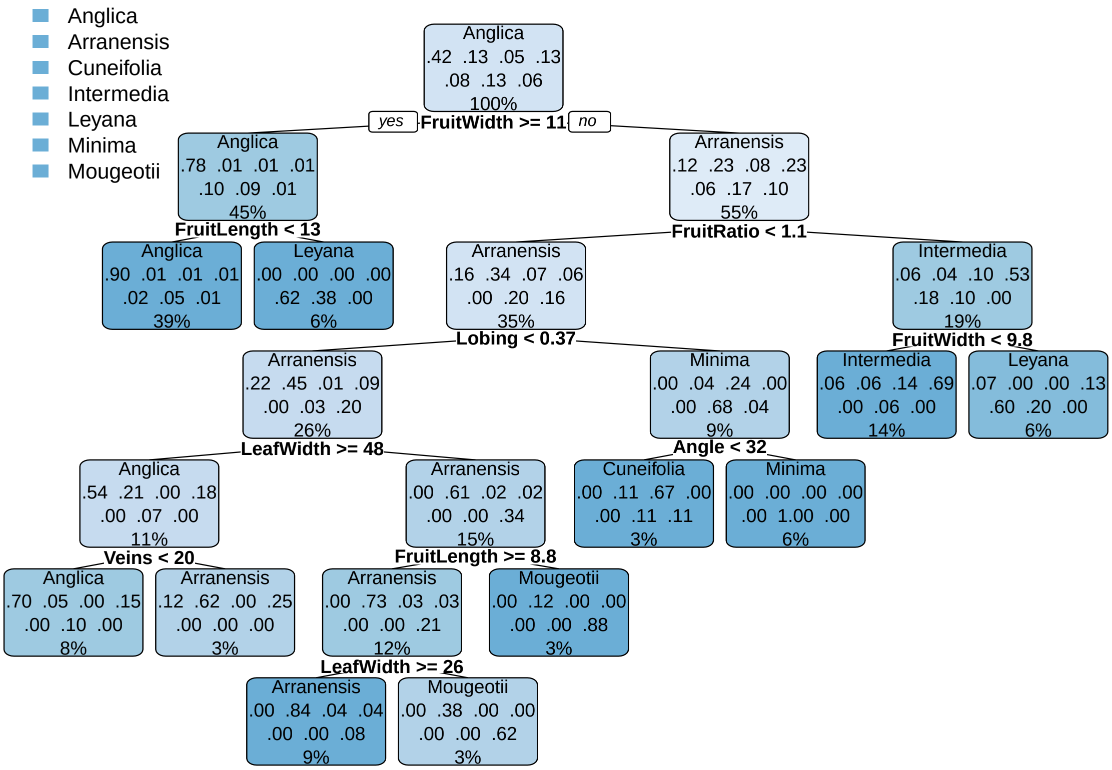
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Anglica .42 .13 .05 .13 .08 .13 .06 100%
yes
FruitWidth >= 11
no
Anglica .78 .01 .01 .01 .10 .09 .01 45%
Arranensis .12 .23 .08 .23 .06 .17 .10 55%
FruitLength < 13
FruitRatio < 1.1
Anglica .90 .01 .01 .01 .02 .05 .01 39%
Leyana .00 .00 .00 .00 .62 .38 .00 6%
Arranensis .16 .34 .07 .06 .00 .20 .16 35%
Intermedia .06 .04 .10 .53 .18 .10 .00 19%
Lobing < 0.37
FruitWidth < 9.8
Arranensis .22 .45 .01 .09 .00 .03 .20 26%
Minima .00 .04 .24 .00 .00 .68 .04 9%
Intermedia .06 .06 .14 .69 .00 .06 .00 14%
Leyana .07 .00 .00 .13 .60 .20 .00 6%
LeafWidth >= 48
Angle < 32
Anglica .54 .21 .00 .18 .00 .07 .00 11%
Arranensis .00 .61 .02 .02 .00 .00 .34 15%
Cuneifolia .00 .11 .67 .00 .00 .11 .11 3%
Minima .00 .00 .00 .00 .00 1.00 .00 6%
Veins < 20
FruitLength >= 8.8
Anglica .70 .05 .00 .15 .00 .10 .00 8%
Arranensis .12 .62 .00 .25 .00 .00 .00 3%
Arranensis .00 .73 .03 .03 .00 .00 .21 12%
Mougeotii .00 .12 .00 .00 .00 .00 .88 3%
LeafWidth >= 26
Arranensis .00 .84 .04 .04 .00 .00 .08 9%
Mougeotii .00 .38 .00 .00 .00 .00 .62 3%

Figure 3: Decision tree unstandardized data

The plot shows the decision tree for the unstandardised data. The decisions at the nodes can be seen to be based initially on fruit width, which was seen in the boxplots to be an variable which differentiate S anglica. The leaves of the tree show that many classes are very accurate. S. minima is 100% accurate, S. anglica 90%. S mougeotii and S arranensis are both about 80% and the remaining species are above 60%.

```
The accuracy for the decision tree is
```

```
0.68
```

Table 15: Sensitivity for the decision tree

Species	Sensitivity
Anglica	0.85
Arranensis	0.47
Cuneifolia	0.17
Intermedia	0.71
Leyana	0.67
Minima	0.6
Mougeotii	0.43

The sensitivity for the decision tree is only 0.17 for S cuneifolia but 0.85 for S anglica.

Table 16: Precision for the decision tree

Class	Precision
class_Anglica	0.79
class_Arranensis	0.5
class_Cuneifolia	0.5
class_Intermedia	0.67
class_Leyana	0.4
class_Minima	1
class_Mougeotii	0.43

The decision tree achieves high precision.

A summary of precision and sensitivity for hierarchical clustering and the decision tree are shown below. The kmeans has not been included since the inconsistency of the method demonstrates that it is not suitable for this data.

Table 17: Confusion matrix for the decision tree

	Anglica	Arranensis	Cuneifolia	Intermedia	Leyana	Minima	Mougeotii	
Anglica	41	3	0	0	3	0	1	48
Arranensis	4	7	0	0	1	0	3	15
Cuneifolia	0	1	1	4	0	0	0	6
Intermedia	2	0	1	10	1	0	0	14
Leyana	3	0	0	0	6	0	0	9
Minima	1	0	0	1	4	9	0	15
Mougeotii	1	3	0	0	0	0	3	7

Table 18: Sensitivity for hierarchical clustering and decision tree

	hclust standardised	hclust non-standardized	tree
Anglica	0.68	0.32	0.85
Cuneifolia	0.28	0.04	0.47
Intermedia	0	0	0.17
Leyana	0.46	0.21	0.71
Minima	0	0.27	0.67
Mougeotii	0.06	0.04	0.6
Arranensis	0.48	0.87	0.43

Table 19: Precision for hierarchical clustering and decision tree

	hclust standardised	hclust non-standardized	tree
Class1	0.69	0.81	0.79
Class2	0.08	0.17	0.5
Class3	0	0	0.5
Class4	0.04	0.37	0.67
Class5	0.05	0	0.4
Class6	0.05	0.43	1
Class7	0.58	0.61	0.43

6 Conlcusion.

As expected, kmeans was not successful in separating the data into clusters which could be interpreted as species of Sorbus. The kmeans algorithm was seen to be unrepeatable and although the ratio of the within cluster sum of squares to between cluster sum of squares was low, the accuracy was always less 0.3. Sometimes high precision or sensitivity was achieved for a single species, but this was not reflected in the other species and it was not repeatable. The standardized data gave slighty better results, as expected.

Hierarchical clustering achieved an accuracy of 0.42 using the Canberra method in standardized data and 0.41 using the Euclidean and Minowski method in non-standardized data. The confusion matrix for the non-standardized data showed better allocation of S Anglica but the confusion matrix for standardised data was better for allocating S mougeotii. The sensitivity and precision also gave inconsistent results for the standardized and non-standardized data. Neither data treatment being better overall for all species.

The decision tree method performed more consistently that hierarchical clustering. Although a single species might have a higher sensitivity in clustering, across all species the decision tree performed better, with five of the seven species achieving greater than 0.6 sensitivity and precision above 0.5 in all but one class. The overall accuracy was also the highest at 0.68.

The tree plot has the added benefit of providing a decision that can be used to assist biological recording.

7 Further work

All the variables were used in the decision tree, which may not be the best model. Rpart provides information on the importance of variables which could be used to ascertain which variables could be removed. The model should be repeated with other species of Sorbus to see if the success was due to the specific morphological characteristics of the Soraria subgenus. It would also be interesting to examine the within plant and within species variability in order to ascertain whether the noise in the data can be reduced or has some minimum level which will always be present.