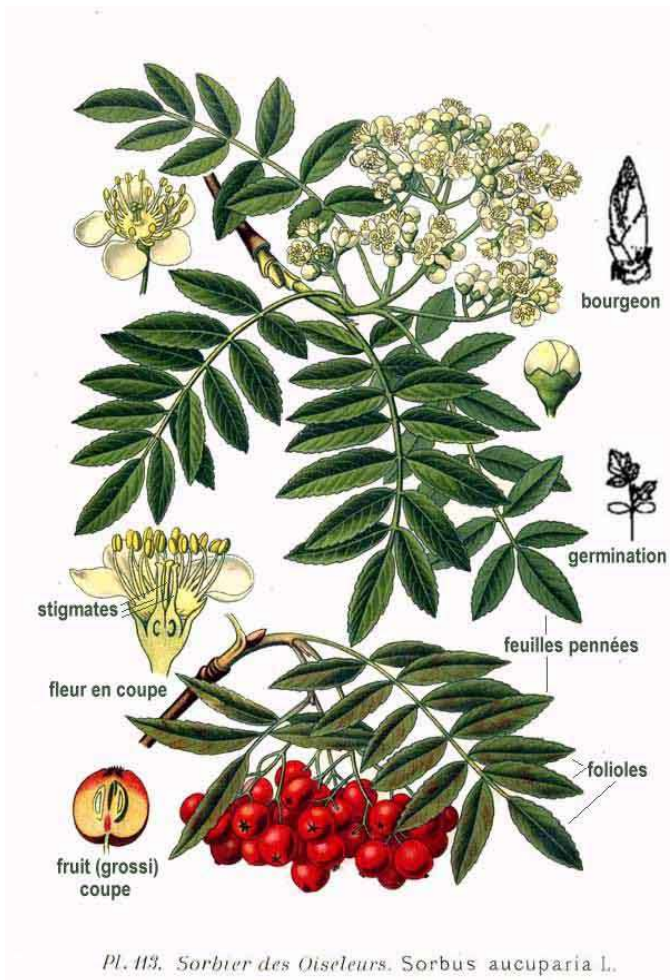
Can machine Learning be used to identify species of Sorbus

PetraGuy, Imperial College London

February 24 2018



0. Abstract.

Three algorithms were used to separate seven species of Sorbus within the subgenus Soraria based on morphological measurements of fruit and leaves. Two unsupervised clustering techniques, namely kmeans and hierarchical clustering and one supervised decision tree. Box plots showed considerable overlap of characteristics between the species, but that some species were differentiated in one or two characteristics, suggesting clustering techniques would be less successful than decision tree methods. This was seen in the results with kmeans modelling being unstable and unable to repeatedly produce accurate results. Hierarchical clustering using the canbera distance metric gave an accuracy of 0.42 while precision ranged from 0 to 0.81 and sensitivity from 0 to 0.87 using non-standardised data. A decision tree was the most successful method giving an accuracy of 0.68 with precision ranging from 0.4 to 1 and sensitivity from 0.17 to 0.85.

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, ("Https://Species.nbnatlas.org/Search/?q=Sorbus&fq=," n.d.), 38 of these are vulnerable or critically endangered and most are endemic or native (H. Rich T. 2010). There are four diploid species, but, as with many Rosaceae, Sorbus produce new apomictic polyploid species, (Robertson 2016). These can also produce viable pollen and can therefore backcross with other diploid or polyploid species,(Ludwig 2013). 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, (J. Rich T. 1998). 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 et al, (H. Rich T. 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. Not every plant in each species had a complete set of measurements or the same number of records. For example, S intermedia had 126 observations but S leyana only had 39. This is due to the relative occurrence of the two species. 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. For S. intermedia, for 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, which would be detrimental to the accuracy of the model, (Peters 2005). Initial data exploration did find this reduced the accuracy of K-means.

Therefore, the sparsity was handled in two ways. Firstly, 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. Secondly, for some species, 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 project focusses on modelling techniques and an initial exploration of machine learning methods; it is not intended 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. However, if these plants are phenotypically very plastic, these assumptions may be invalid. The range of leaf sizes within each plant was not available, so a comparison of the variation within each plant and between all the plants of each species would be useful here.

This data handling 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. Clustering algorithms, because they rely on distance metrics, are usually sensitive to scale in the data,(Ismail 2013), therefore the data was also standardized and each clustering model carried out on both standardized and non-standardized data. Since standardization should not effect a decision tree (Nisbet 2017), only the non-standardized data was modelled.

Because each species had a different number of records a random stratified sampling system was carried out to create train and test sets used in the supervised leaning algorithms. Each species was split into 70/30 train/test sets.

3. Modelling

3.1. Model performance metrics

In supervised learning, the correct and incorrect values assigned to each class are known, and these are used to evaluate the model by calculating accuracy, precision and sensitivity.

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

Precision true positive rate of a predicted class. The precision for a species tells you how accurately the algorithm is identifying a species, a low precision tells you that other species are incorrectly lumped with the correct species. A high precision tells you that most of the species are correctly identified and that the predicted class will be predominantly made up of the right species.

Sensitivity is the true positive rate of a species. A low number tells you the correct species have been put in other, incorrect, classes. A high sensitivity tells you that most of the species have been put in the right class, and that most of the actual species are in the correct predicted class.

Actually, despite using a mixture of unsupervised and supervised learning methods, we do know the identification of the species, so in fact we can calculate accuracy, precision and sensitivity for all the models and compare them using the same metric.

In addition, clustering algorithms can use various metrics, such as the ratio of within cluster sum of squares to total sum of squares to evaluate the model. For well defined, compact clusters the ratio will be small. Since this metric was not available for all models, it is not used to compare different models. In addition, since the two clustering techniques performed so poorly, there was no reason to compare the two, and therefore these metrics are not shown here.

Confusion matrices, which summarise the the frequencies of the species allocated to different classes and clusters, were produced to examine two of the models, but they were unfeasible for the K-means algorithms since this was repeated ten times, as discussed below, and the large number of confusion matrices would obfuscate the results. Since they give the same information as accuracy, precision and sensitivity, they were used to discuss hierarchical clustering and the decision tree, but not to compare models or examine the results of K-means.

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. Splits 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 rarely be the case due to noise within the data. The model here is represented by the logical processes followed to reach the final classes. The rpart package was used for the decision tree, (Therneau 2018). The rpart library also offers a decision tree plot which summarises the binary choices used at each node.

3.2.2. K-means

K-means 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 are not specified. In clustering, no information is needed about the objects and there is no right or wrong, so in that sense, our problem does not demand clustering. We know what species a sample belongs to and we do not want it allocated to another cluster. However, it is a useful technique to see if the model reflects the patterns we know the data contains. The K in K-means refers to the number of clusters to be used, which, because we know the data contains seven species we specified as such.

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, 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 accuracy on each repeat will indicate the success of the model. Ten repeats of the model were carried out and the accuracy calculated on each run.

3.2.3. Hierarchical Clustering.

Bottom up 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. The clusters can be cut at k = 7 and the members can be examined. Hierarchical clustering was explored using different distance methods in order to ascertain the method giving the highest accuracy and this method was then used to calculate precision and sensitivity. The hclust function was used which 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. 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.

If the data is separated into clearly defined groups, we can be sure that a clustering algorithm will work. Box plots are presented for the standardized and non-standardized data and show how the scale and separation between groups could be an issue for the clustering algorithms. The plots also show 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.

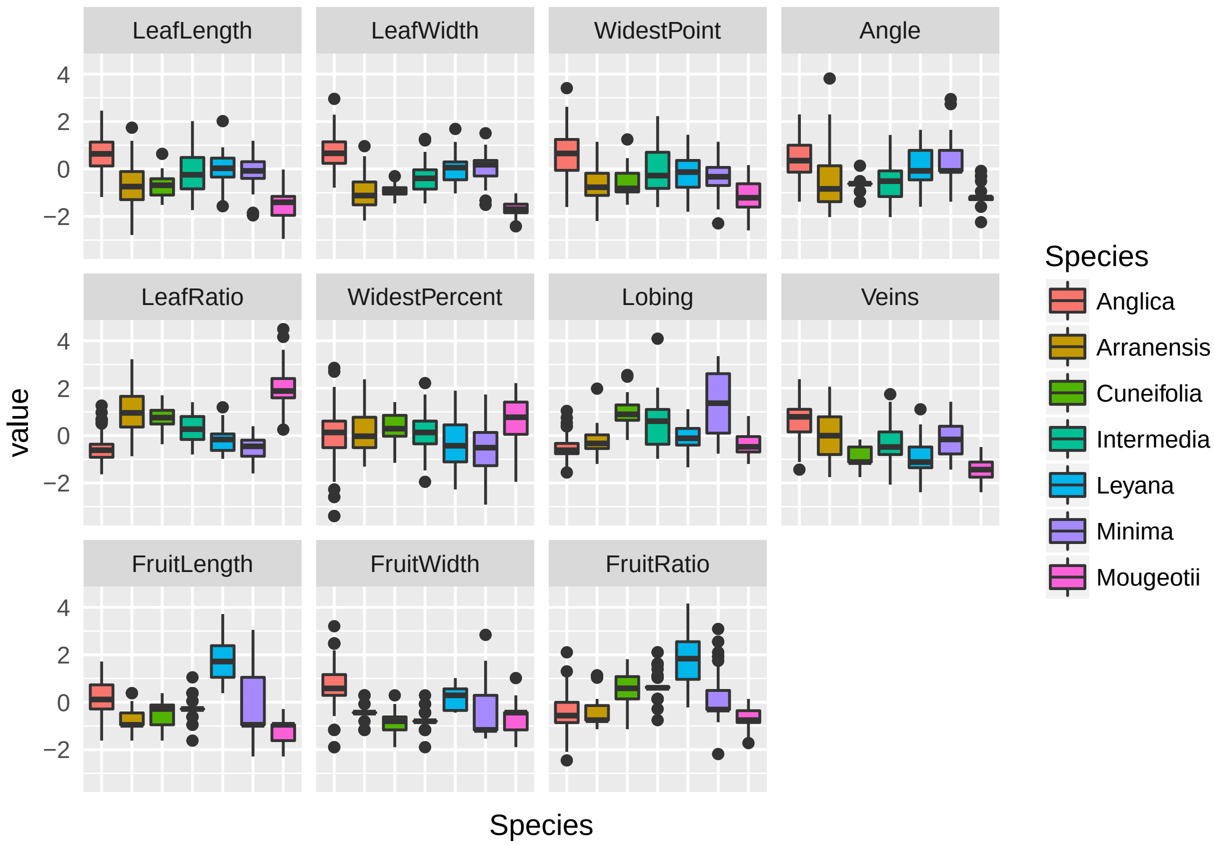


Figure 1: Box plots for standardized data

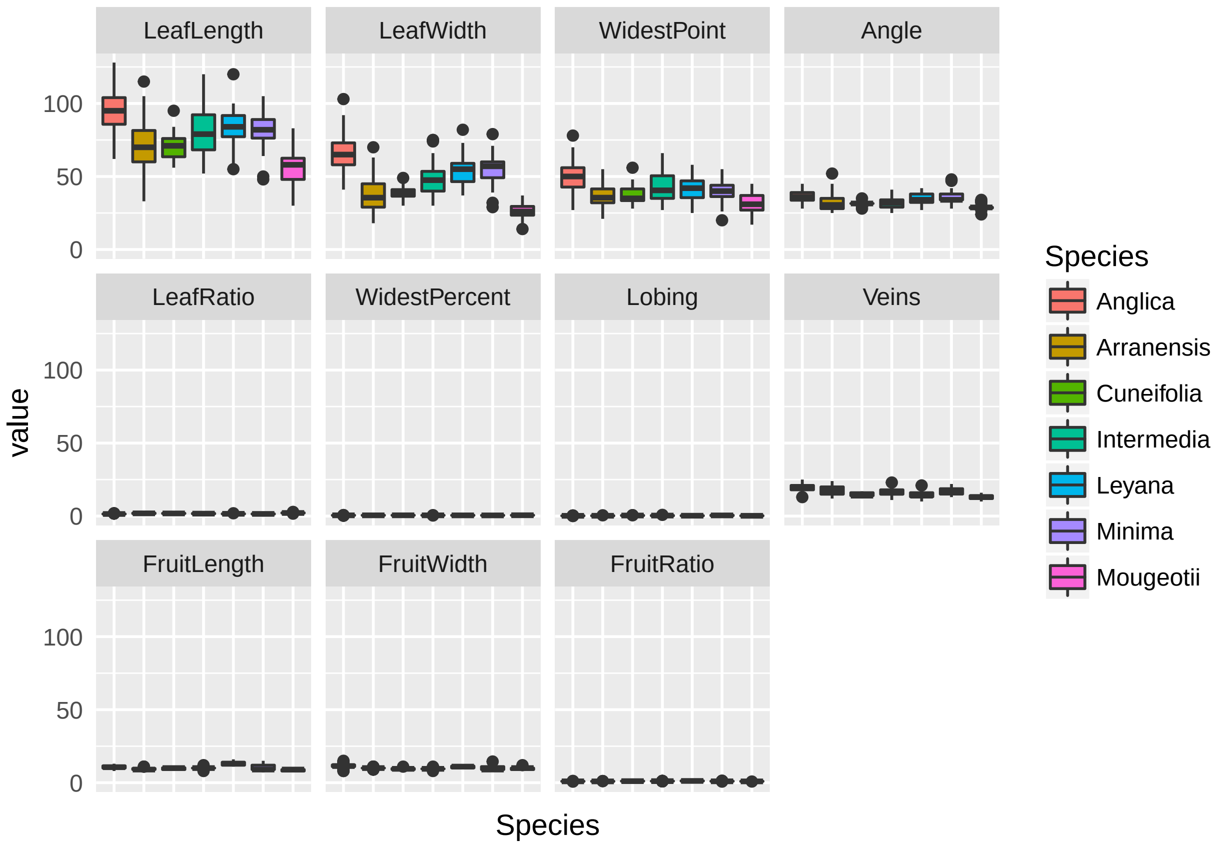


Figure 2: Box plots for non-standardized data

5. Results

5.1 Kmeans

The accuracy shown in table 1 of the appendix is different on each run implying that the algorithm is not successfully grouping the data into the same clusters.

Table 2 shows 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 K-means algorithm.

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. Tables 4 to 7 show that 100% sensitivity and precision could be achieved on the standardized data for S. arranensis, but this was not seen in other species and in a subsequent run this would drop to 0.

In summary, K-means is not consistent across species, does not achieve high accuracy and is not repeatable.

5.2 Hierarchical Clustering

Tables 8 and 9 show that the Canberra metric gives an accuracy of 0.42 for the non-standardised data. The Euclidean and Minkowski methods give slightly worse accuracy of 0.41 for the standardized data, as in K-mean, this is contrary to expectations. The sensitivity and precision calculated for standardised and non-standardised data are shown in tables 10 to 13.

The results are again inconsistent, a high precision and sensitivity of 0.69 and 0.68 is achieved for S. anglia using standardized data, but those values are 0 for S. Intermedia.

In summary, the hierarchical clustering technique gives low accuracy and inconsistent precision and sensitivity across the species and classes.

5.3 Decision Tree

Table 14 shows numbers of species in each group, it is useful to be aware of the proportion of species in the test set when analysing the tree plot.

The first decision splits the data roughly in half depending on the fruit being either greater or less than 11mm wide. The thinner fruit is then predominantly assigned to S. anglica based on the length being less than 13mm. 39% of the test set is in this leaf, S.anglica comprises 42% of the test set, this allocation is therefore very accurate. The wider fruits take more decisions to assign the species. Fruit ratio (fruit width/fruit length) and fruit length < 9.8mm gives 14% of the data, most of which is allocated to S.intermedia with some S. cuneifolia. The S. minima leaf contains only that species (high precision) but only 6% of the data instead of 13% so we can see that around half this species has been incorrectly assigned (low sensitivity).

The accuracy of the tree, shown in table 15, 0.68. The confusion matrix in table 16 shows that most of S.anglica, S. intermedia, S.leyana, and S.minima are grouped together. The sensitivity, shown in table 17, is above 43% for 6 out of the 7 species while table 18 shows that the precision is above 50% for 6 classes and 100% for S.minima.

A summary of precision and sensitivity for hierarchical clustering and the decision tree are shown in tables 19 and 20. The K-means has not been included since the inconsistency of the method demonstrates that it is not suitable for this data.

In summary, the decision tree is the most successful of the algorithms achieving the highest accuracy of 0.68, with precision and sensitivity being consistently higher across the classes and species.

6 Conclusion.

K-means was not successful in separating the data into clusters which could be interpreted as species of Sorbus. The algorithm was seen to be unrepeatable and 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 slightly better results. It is not clear from this analysis whether it is the nature of the data itself that is the issue; Raykov et al (Raykov 2016) describe the need for datasets of equal variance and size, which was not the case here, or the data preparation. Different methods standardisation have been shown to influence the outcome of K-means, and that the method used here may not be the optimum, (Banks 2011).

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

Different methods of standardisation could be tried for the unsupervised methods, as well as other clustering algorithms which may be better able to model this data, (**???**).

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 could be removed.

The model could be extended to include cross fold validation, which was not incorporated in the supervised learning algorithm. Other species of Sorbus could be modelled 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.

Appendix I K-means results.

Table 1: Accuracy

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
unstandardized	0.10	0.08	0.21	0.18	0.10	0.18	0.08	0.16	0.09	0.09
standardized	0.02	0.02	0.20	0.08	0.04	0.27	0.13	0.03	0.01	0.02

Table 2: Percentage of true positives for non-standarized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0.00	0.00	18.75	18.75	13.12	24.38	0.00	25.00	3.12	0.00
Cuneifolia	34.00	20.00	20.00	30.00	24.00	4.00	6.00	34.00	24.00	34.00
Intermedia	5.26	57.89	0.00	21.05	5.26	21.05	5.26	0.00	5.26	21.05
Leyana	4.17	2.08	31.25	31.25	2.08	4.17	31.25	2.08	2.08	2.08
Minima	3.33	23.33	23.33	3.33	3.33	3.33	36.67	3.33	3.33	23.33
Mougeotii	32.00	4.00	4.00	8.00	6.00	30.00	4.00	4.00	28.00	8.00
Arranensis	0.00	0.00	73.91	0.00	0.00	21.74	0.00	4.35	0.00	0.00

Table 3: Percentage of true positives for standarized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0.00	0.00	26.25	1.88	0.00	49.38	0.00	0.00	0.00	0.00
Cuneifolia	8.00	6.00	0.00	4.00	0.00	0.00	4.00	0.00	2.00	4.00
Intermedia	0.00	5.26	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
Leyana	4.17	4.17	20.83	4.17	18.75	0.00	4.17	12.50	4.17	4.17
Minima	6.67	0.00	83.33	6.67	6.67	0.00	0.00	6.67	6.67	0.00
Mougeotii	0.00	0.00	0.00	46.00	10.00	6.00	46.00	6.00	0.00	6.00
Arranensis	0.00	0.00	0.00	0.00	0.00	0.00	100.00	8.70	0.00	0.00

Table 4: Precision of kmeans with non standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0	0	0.19	0.19	0.13	0.24	0	0.25	0.03	0
Cuneifolia	0.34	0.2	0.2	0.3	0.24	0.04	0.06	0.34	0.24	0.34
Intermedia	0.05	0.58	0	0.21	0.05	0.21	0.05	0	0.05	0.21
Leyana	0.04	0.02	0.31	0.31	0.02	0.04	0.31	0.02	0.02	0.02
Minima	0.03	0.23	0.23	0.03	0.03	0.03	0.37	0.03	0.03	0.23
Mougeotii	0.32	0.04	0.04	0.08	0.06	0.3	0.04	0.04	0.28	0.08
Arranensis	0	0	0.74	0	0	0.22	0	0.04	0	0

Table 5: Precision of kmeans with standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0	0	0.26	0.02	0	0.49	0	0	0	0
Cuneifolia	0.08	0.06	0	0.04	0	0	0.04	0	0.02	0.04
Intermedia	0	0.05	0	0	0	1	0	0	0	0
Leyana	0.04	0.04	0.21	0.04	0.19	0	0.04	0.12	0.04	0.04
Minima	0.07	0	0.83	0.07	0.07	0	0	0.07	0.07	0
Mougeotii	0	0	0	0.46	0.1	0.06	0.46	0.06	0	0.06
Arranensis	0	0	0	0	0	0	1	0.09	0	0

Table 6: Sensitivity of kmeans with non standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0	0	0.19	0.19	0.13	0.24	0	0.25	0.03	0
Cuneifolia	0.34	0.2	0.2	0.3	0.24	0.04	0.06	0.34	0.24	0.34
Intermedia	0.05	0.58	0	0.21	0.05	0.21	0.05	0	0.05	0.21
Leyana	0.04	0.02	0.31	0.31	0.02	0.04	0.31	0.02	0.02	0.02
Minima	0.03	0.23	0.23	0.03	0.03	0.03	0.37	0.03	0.03	0.23
Mougeotii	0.32	0.04	0.04	0.08	0.06	0.3	0.04	0.04	0.28	0.08
Arranensis	0	0	0.74	0	0	0.22	0	0.04	0	0

Table 7: Sensitivity of kmeans with standardized data

	1	2	3	4	5	6	7	8	9	10
Anglica	0	0	0.26	0.02	0	0.49	0	0	0	0
Cuneifolia	0.08	0.06	0	0.04	0	0	0.04	0	0.02	0.04
Intermedia	0	0.05	0	0	0	1	0	0	0	0
Leyana	0.04	0.04	0.21	0.04	0.19	0	0.04	0.12	0.04	0.04
Minima	0.07	0	0.83	0.07	0.07	0	0	0.07	0.07	0
Mougeotii	0	0	0	0.46	0.1	0.06	0.46	0.06	0	0.06
Arranensis	0	0	0	0	0	0	1	0.09	0	0

Appendix II Hierachical clustering results

Table 8: 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 9: 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

Table 10: Confusion matrix for Canberra method using non-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

Table 11: 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

Table 12: 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

Table 13: 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

Appendix 3 Decision tree resuilts

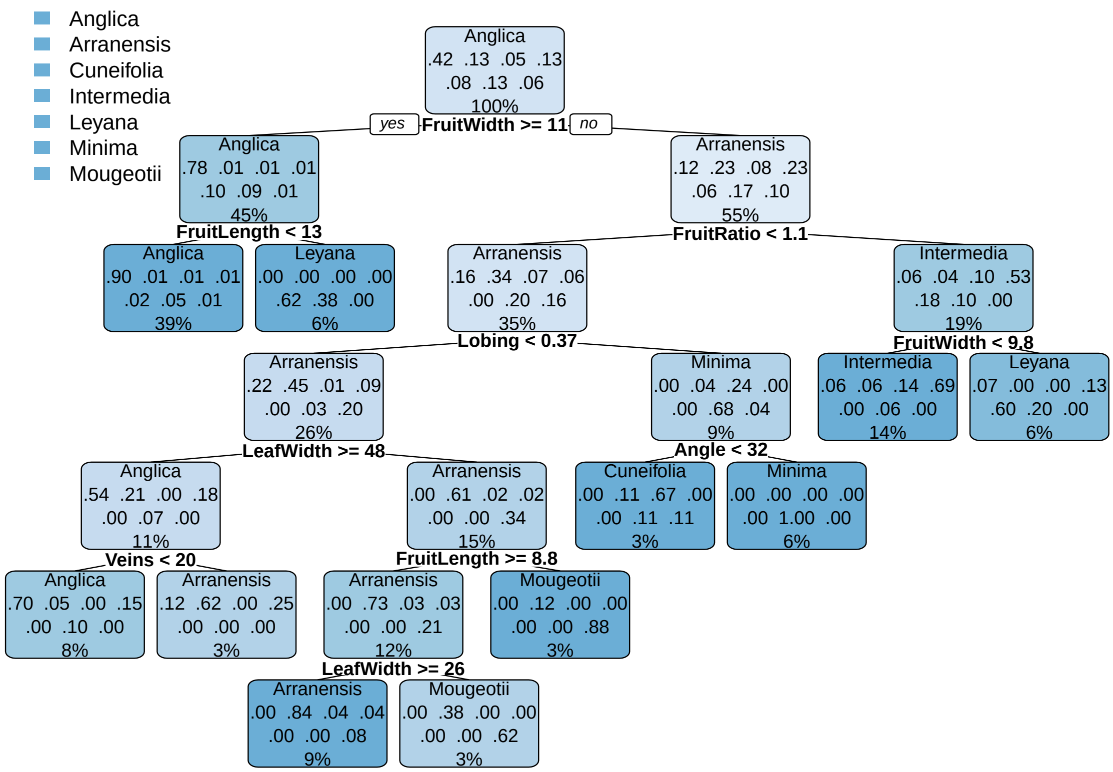


Figure 3: Decision tree unstandardized data

Table 14: Proportions of each species in test set

Anglica	0.42
Arranensis	0.13
Cuneifolia	0.05
Intermedia	0.13
Leyana	0.08
Minima	0.13
Mougeotii	0.06

Table 15: Accuracy of the decision tree

0.68

Table 16: 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 17: 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

Table 18: 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

Table 19: 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 20: 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

References

Banks, House, D., ed. 2011. *Standardizing Variables in K-Means Clustering.* Springer Science; Business Media.

"Https://Species.nbnatlas.org/Search/?q=Sorbus&fq=." n.d.

Ismail, Dauda, B.M. 2013. "Standardization and Its Effect on K-Means Clustering Algorithm." *Research Journal of Applied Sciences Engineering and Technology* 6 (17). Maxwell scientific: 3299–3303.

Ludwig, S. 2013. "Breeding Systems, Hybridization and Continuing in Avon Gorge Sorbus." *Anals of Botany* 111 (4). Oxford academic: 563–75.

Nisbet, Miner, R. 2017. *Handbook on Statistical Analysis and Data Mining Applications.* Elsevier.

Peters, F., J. 2005. *Transactions on Rough Sets Iv.* Springer Science; Business Media.

Raykov, Boukouvalal, Y.P. 2016. "What to Do When K-Means Clustering Fails: A Simple Yt Principled Alternative Algorithm." *PLoSONE* 11 (9). https://doi.org/10.137/journal.pone0162259: 376–94.

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

Rich, Jermy, T. 1998. *Plant Crib.* BSBI.

Robertson, Phipps, K.R. 2016. "A Synopsis of Genera in Maloideae(Rosaceae)." *Research Journal of Applied Sciences Engineering and Technology* 16 (2). American Society of Plant Taxonomists: 376–94.

Therneau, Atkinson, T. 2018. *Recursive Partitioning and Regression Trees.* Vienna, Austria: R foundation for statistical computing.