A Classification of Plant Species with the Use of Image Processing Results from the Leaf Dataset

Written by Tommy Doody and Cole Zarifis

April 9, 2021

Data Science II
MTH-3510-01
Professor Petkus
Aurora University

Executive Summary

Six methods of classification were used on three different versions of the Leaf Dataset. These methods include K-Nearest Neighbor, Linear Discriminant Analysis, Classification Tree, Bagging, Random Forests and Support Vector Machines. The first version of the dataset used all variables provided, the second used all variables with the inclusion of four interaction variables, and the last using only significant variables. The LDA model on the interaction dataset performed the best with an initial accuracy of 80% on the test set, with cross validation accuracy of 78.08% of one-thousand iterations.

Introduction

The Leaf Dataset is being examined with the intention of accurately classifying the variable Species using various classification and clustering methods. The Leaf Dataset contains a total of 340 observations of leaf specimens obtained from photographs taken by an Apple iPAD 2. The 24-bit RGB images recorded have a resolution of 720 x 920 pixels. The Species variable that is being evaluated has 40 distinct data points corresponding to different plant species including Quercus Suber, Salix Atrocinera, and Quercus Rober. However, only 30 of the 40 species are recorded in the Leaf Dataset. For each of the data points, information is given regarding the shape and texture of the specimens from the photographs. The purpose of this analysis is to identify if classifying plant species using images is possible and to evaluate the effectiveness of this approach.

Analysis

Dataset Cleaning, Classification Models, and Assumptions

The Leaf Dataset has a total of sixteen variables: Species, Specimen Number, Eccentricity, Aspect Ratio, Elongation, Solidity, Stochastic Convexity, Isoperimetric Factor, Maximal Indentation Depth, Lobedness, Average Intensity, Average Contrast, Smoothness, Third Moment, Uniformity, and Entropy. The first constructed classification model will utilize the entire dataset to classify Species. Although, Specimen Number will be removed from the Dataset since it does not provide any insightful information in regards to image processing and the classification of Species. By observing the chart correlation (Figure 1), it was noted that none of the variables appear to have a high correlation with Species. However, the variables Aspect Ratio, Eccentricity, Elongation, Solidity, Average Intensity, Smoothness, and Uniformity have relatively high significance with Species. In light of this finding, a second classification model was constructed to observe how the significant variables classify Species. Furthermore, from the chart correlation, there appeared to be three groups of variables that were highly correlated with members of the group, but not with other groups. The variable that had the highest correlations with all other members of the group was used for adding interactions. Eccentricity, Solidity and Average Intensity were picked and 4 interactions were performed using the possible combinations of the three variables. For test/train purposes, 250 of the data points will be used for training (73.53% of the observations) and 90 will be used for testing (26.47% of the observations).

K-Nearest Neighbors Method

For the K-Nearest Neighbors (KNN) method, it was decided that the first 41 values of k would be observed. No value of k was pursued beyond 41 due to computational limitations. One thousand simulations were performed using the KNN method on random samples from the Leaf Dataset for each classification model. As can be seen by the results (Figure 2), the best value for k in each instance was k = 1 with a success rate of 60.00% for the Whole Dataset model, 55.50% for the Significant Variables model, and 58.89% for the Interactions model.

Linear Discriminant Analysis Method

One thousand simulations were performed using the Linear Discriminant Analysis (LDA) method on random samples from the Leaf Dataset for each classification model. From the results of these simulations, it was found that the success rates in each instance were 76.67% for the Whole Dataset model, 66.24% for the Significant Variables model, and 78.08% for the Interactions model.

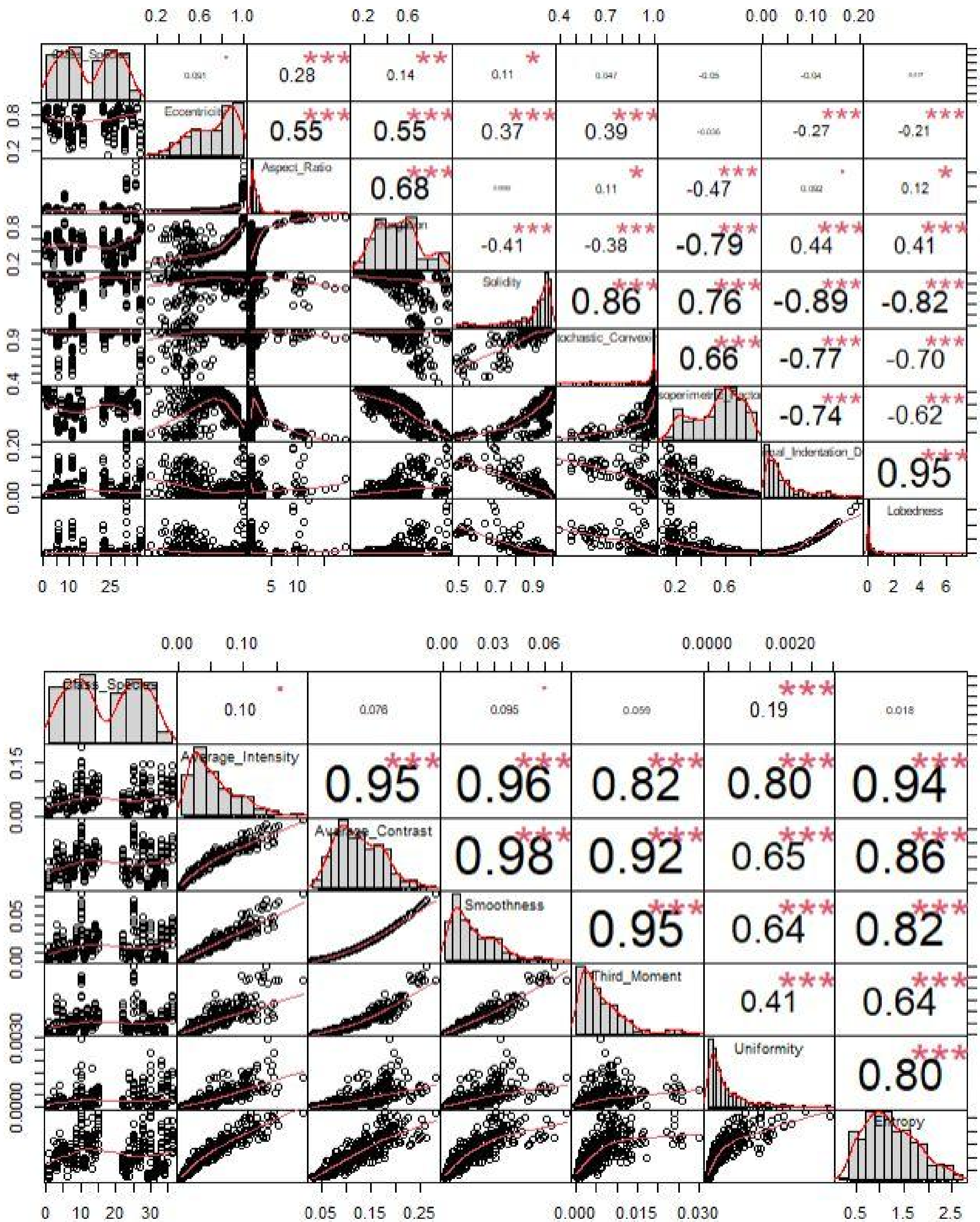

Figure 1: Chart Correlations of the Leaf Dataset

	Whole Dataset	***Significant Variables***	***Interactions***
1	0.6	0.5549667	0.5888889
3	0.6	0.4688667	0.5777778
5	0.5111111	0.4593444	0.5111111
7	0.5444444	0.4343111	0.5777778
9	0.5333333	0.4088778	0.5777778
11	0.5777778	0.3882444	0.5888889
13	0.5444444	0.3700111	0.5333333
15	0.5555556	0.3507667	0.5222222
17	0.5111111	0.3321	0.5
19	0.4777778	0.3163889	0.4777778
21	0.4555556	0.301	0.4555556
23	0.4222222	0.2883222	0.4222222
25	0.4111111	0.2764667	0.3666667
27	0.3666667	0.2655889	0.3777778
29	0.3666667	0.2570778	0.3444444
31	0.3444444	0.2476222	0.3777778
33	0.3777778	0.2388333	0.3333333
35	0.3111111	0.2298667	0.3555556
37	0.3222222	0.2206778	0.3444444
39	0.3555556	0.2134556	0.3333333
41	0.3333333	0.2075333	0.2888889

Figure 2: Results of the KNN method for each Classification Model

Classification Tree Method

When constructing the classification trees, it was noted that ten of the fourteen variables from the Leaf Dataset were used in the Whole Dataset tree, all significant variables were used in the Significant Variable tree, and ten variables were used in the Interactions tree. Of the ten variables used in the Interactions tree, only one of them was an interaction (the interaction between Eccentricity and Solidity). One thousand simulations were performed using the Classification Tree method on random samples from the Leaf Dataset for each classification model. From the results of these simulations, it was found that the success rates in each instance were 45.56% for the Whole Dataset model, 43.55% for the Significant Variables model, and 44.44% for the Interactions model. In terms of pruning, it was determined that pruning is best at a size of 25 for the Whole Dataset model, 26 for the Significant Variable model, and 24 for the Interactions model (Figure 3). Which is to say, none of the models benefit from pruning and is therefore considered unnecessary.

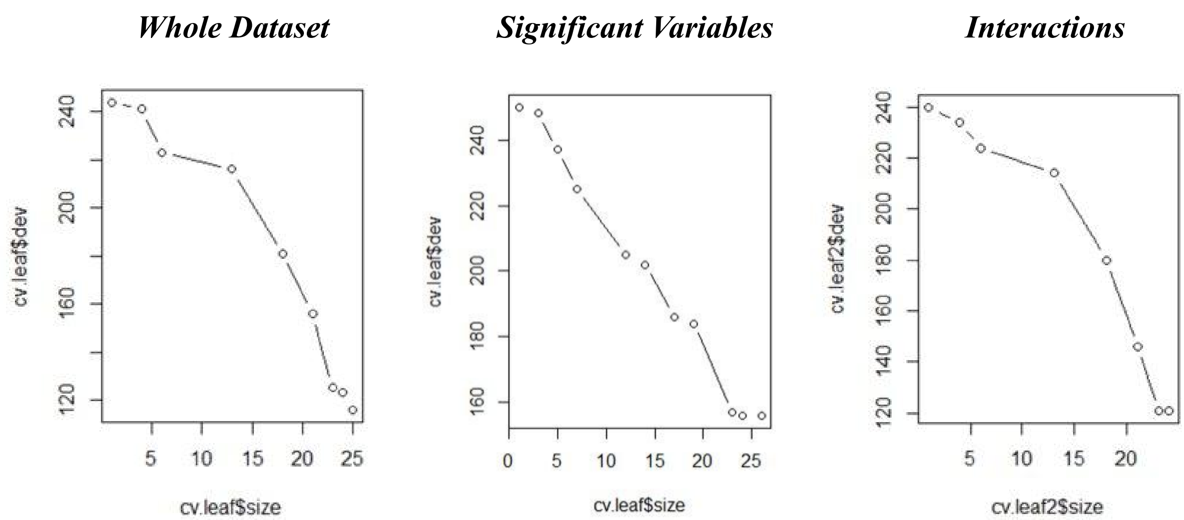

Figure 3: Size vs. Deviation Charts for each Classification Model

Bagging Method

One thousand simulations were performed using the Bagging method on random samples from the Leaf Dataset for each classification model. From the results of these simulations, it was found that the success rates in each instance were 72.22% for the Whole Dataset model, 67.19% for the Significant Variables model, and 74.44% for the Interactions model.

Random Forests Method

For the random forests method, the number of variables tried at each split are determined by the value of the argument mtry. This argument was tested at every possible value in each of the classification models to observe the success rate in each instance (Figure 4). One thousand simulations were performed using the Random Forests method on random samples from the Leaf Dataset for each classification model. From the results of these simulations, it was found that the best success rates in each instance were 77.78% at mtry = 2 for the Whole Dataset model, 68.08% at mtry = 3 for the Significant Variables model, and 77.78% at mtry = 3 for the Interactions model.

Whole Dataset

```
mtry: 2 accuracy: 0.7777778
mtry: 3 accuracy: 0.7666667
mtry: 4 accuracy: 0.7777778
mtry: 5 accuracy: 0.7666667
mtry: 6 accuracy: 0.7666667
mtry: 7 accuracy: 0.7666667
mtry: 8 accuracy: 0.7666667
mtry: 9 accuracy: 0.7555556
mtry: 10 accuracy: 0.7555556
mtry: 11 accuracy: 0.7444444
mtry: 12 accuracy: 0.7444444
mtry: 13 accuracy: 0.7333333
```

Interactions

```
mtry: 2 accuracy: 0.7666667
mtry: 3 accuracy: 0.7777778
mtry: 4 accuracy: 0.7555556
mtry: 5 accuracy: 0.7666667
mtry: 6 accuracy: 0.7666667
mtry: 7 accuracy: 0.7555556
mtry: 8 accuracy: 0.7555556
mtry: 9 accuracy: 0.7444444
mtry: 10 accuracy: 0.7444444
mtry: 11 accuracy: 0.7444444
mtry: 12 accuracy: 0.7555556
mtry: 13 accuracy: 0.7555556
mtry: 14 accuracy: 0.7555556
mtry: 15 accuracy: 0.7555556
mtry: 16 accuracy: 0.7555556
mtry: 17 accuracy: 0.7444444
```

Significant Variables

```
       V1              V2              V3              V4              V5              V6
Min.   :0.5960   Min.   :0.5880   Min.   :0.5960   Min.   :0.5880   Min.   :0.6000   Min.   :0.5960
1st Qu.:0.6440   1st Qu.:0.6640   1st Qu.:0.6640   1st Qu.:0.6640   1st Qu.:0.6600   1st Qu.:0.6560
Median :0.6600   Median :0.6800   Median :0.6800   Median :0.6800   Median :0.6760   Median :0.6720
Mean   :0.6591   Mean   :0.6779   Mean   :0.6808   Mean   :0.6791   Mean   :0.6767   Mean   :0.6741
3rd Qu.:0.6760   3rd Qu.:0.6920   3rd Qu.:0.6960   3rd Qu.:0.6960   3rd Qu.:0.6920   3rd Qu.:0.6880
Max.   :0.7440   Max.   :0.7440   Max.   :0.7560   Max.   :0.7480   Max.   :0.7480   Max.   :0.7520
```

Figure 4: Results of the Random Forests method for each Classification Model

Support Vector Machines Method

Utilizing the best model command to determine the best values for cost, degree, and gamma, one thousand simulations were performed for each support vector machine method (polynomial, linear, and radial) on random samples from the Leaf Dataset for each classification model. From the results of these simulations, it was found that the Whole Dataset model had success rates of 72.22% for the polynomial method with a degree of 2 and a cost of 9.01, 75.56% for the linear method with a cost of 7.91, and 70.00% for the radial method with a gamma of 0.21 and a cost of 9.61. The Significant Variables model had success rates of 55.41% for the polynomial method with a degree of 2 and a cost of 7.767, 68.49% for the linear method with a cost of 7.767, and 64.90% for the radial method with a gamma of 0.51 and a cost of 7.767. The Interactions model had success rates of 73.33% for the polynomial method with a degree of 2 and a cost of 9.41, 70.00% for the linear method with a cost of 2.51, and 68.89% for the radial method with a gamma of 0.21 and a cost of 7.41.

Principal Component Analysis

By performing Principal Component Analysis (PCA) on the Leaf Dataset, it was discovered that one group of variables including Solidity, Stochastic Convexity, and Isoperimetric Factor share similar attributes (Figure 5). Another group that was identified includes Lobedness, Maximal Indentation Depth, and Elongation. The largest group identified includes Uniformity, Smoothness, Third Moment, Solidity, Entropy, and Average Intensity. Separate from these groups exist Eccentricity and Aspect Ratio. These observations reinforce that the variables Eccentricity, Aspect Ratio, Elongation, Solidity, Average Intensity, Smoothness, and Uniformity all have significance in the Leaf Dataset based on their relevance in each of these groupings and influence on the data.

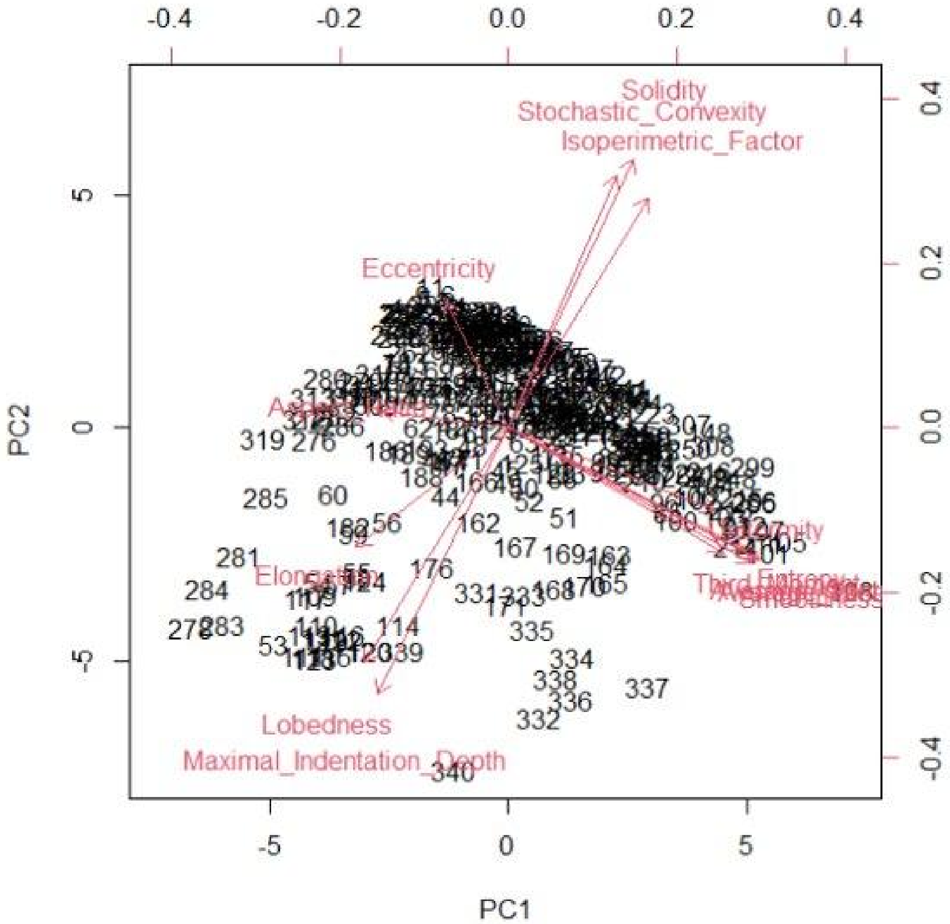

Figure 5: Principal Component Analysis on the Leaf Dataset

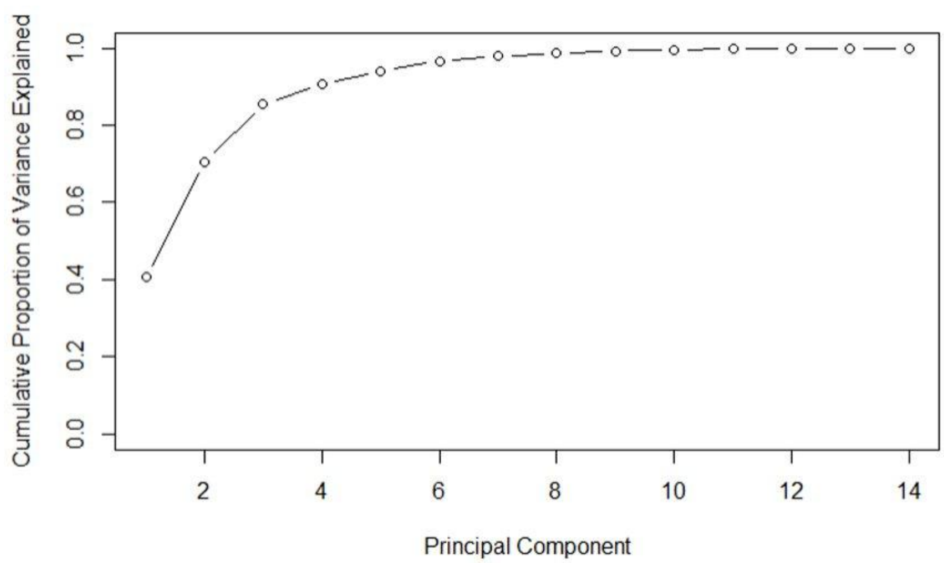

Figure 6: Cumulative Percentage of Variance Explained vs. Principal Components

By plotting the cumulative percentage of variance explained (PVE) of each principal component (Figure 6), it was discovered that 85.57% of the variance is explained when using the first three principal components. This is relatively good and it was decided that the first three principal components should be used in PCA clustering. By performing PCA clustering, it was discovered that certain species were closely related to one another (Figure 7). One group identified was comprised of Species #6 (Crataegus monogyna) and Species #11 (Acer palmatum). Another group was remarked to include Species #8 (Nerium oleander), Species #31 (Podocarpus sp.), and Species #34 (Pseudosasa japonica). The final group identified consisted of Species #10 (Tilia tomentosa), Species #25 (Arisarum vulgare), and Species #30 (Urtica dioica). The other species are indistinguishable since they all overlap one another. These findings reflect that each grouping of leaves has certain distinct features that separate them from other species. However, leaves within these groups may be harder to classify based on images alone since they each hold similar attributes.

K-Means Clustering

By performing K-Means clustering on the Leaf Dataset, it was discovered that the variables Aspect Ratio, Lobedness, and Entropy are all significant since they all have drastically higher ranges of values than any other variable within the dataset (Figure 8). This reinforces the findings found during principal component analysis and from the chart correlation.

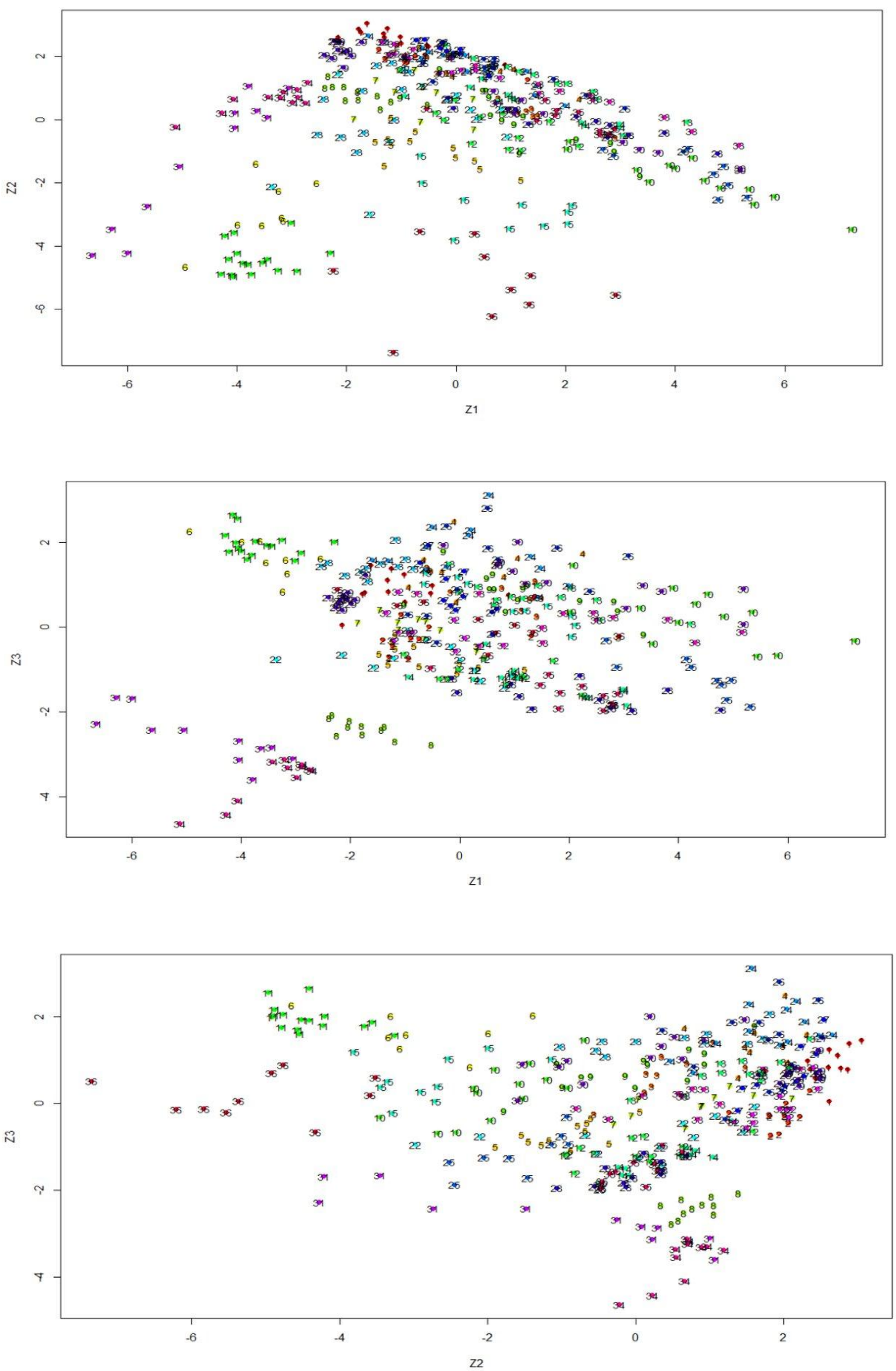

Figure 7: PCA Clustering Results with the First Three Principal Components

```
  Eccentricity     Aspect_Ratio      Elongation        Solidity      Stochastic_Convexity
 Min.   :0.3740   Min.   : 1.091   Min.   :0.2357   Min.   :0.5593   Min.   :0.6128
 1st Qu.:0.5422   1st Qu.: 1.206   1st Qu.:0.4101   1st Qu.:0.8192   1st Qu.:0.9099
 Median :0.8218   Median : 1.758   Median :0.6042   Median :0.9257   Median :0.9836
 Mean   :0.7479   Mean   : 3.525   Mean   :0.5900   Mean   :0.8715   Mean   :0.9229
 3rd Qu.:0.9294   3rd Qu.: 2.886   3rd Qu.:0.6720   3rd Qu.:0.9606   3rd Qu.:0.9946
 Max.   :0.9984   Max.   :16.979   Max.   :0.9421   Max.   :0.9782   Max.   :0.9991
 Isoperimetric_Factor Maximal_Indentation_Depth   Lobedness        Average_Intensity   Average_Contrast
 Min.   :0.1007       Min.   :0.01001          Min.   :0.02397   Min.   :0.007824   Min.   :0.04551
 1st Qu.:0.2173       1st Qu.:0.01956          1st Qu.:0.09084   1st Qu.:0.019228   1st Qu.:0.07509
 Median :0.4844       Median :0.03167          Median :0.22620   Median :0.033031   Median :0.10710
 Mean   :0.4448       Mean   :0.05430          Mean   :1.00383   Mean   :0.046312   Mean   :0.11687
 3rd Qu.:0.6249       3rd Qu.:0.07852          3rd Qu.:1.15197   3rd Qu.:0.065612   3rd Qu.:0.14969
 Max.   :0.7818       Max.   :0.18890          Max.   :6.50610   Max.   :0.134275   Max.   :0.20857
   Smoothness         Third_Moment          Uniformity            Entropy
 Min.   :0.002188   Min.   :0.0006412   Min.   :2.480e-05   Min.   :0.3053
 1st Qu.:0.005799   1st Qu.:0.0019295   1st Qu.:8.006e-05   1st Qu.:0.5811
 Median :0.012040   Median :0.0041842   Median :1.811e-04   Median :0.8683
 Mean   :0.016030   Mean   :0.0054354   Mean   :3.264e-04   Mean   :1.0647
 3rd Qu.:0.022588   3rd Qu.:0.0085490   3rd Qu.:3.655e-04   3rd Qu.:1.5038
 Max.   :0.042457   Max.   :0.0153047   Max.   :1.584e-03   Max.   :2.3823
```

Figure 8: Summary of K-Means Clustering Results for 30 clusters

Hierarchical Clustering

By performing hierarchical clustering on the Leaf Dataset, it was discovered that some species are closely related to one another (Figure 9). One group identified consisted of Species #8 (Nerium oleander), Species #31 (Podocarpus sp.), and Species #34 (Pseudosasa japonica). Another group identified was comprised of Species #6 (Crataegus monogyna) and Species #11 (Acer palmatum). These findings reflect the same discoveries found in PCA clustering. As can be seen from the splits in the trees, each grouping of leaves has certain distinct features that separate them from other species. However, leaves within these groups may be harder to classify based on images alone since they each hold similar attributes.

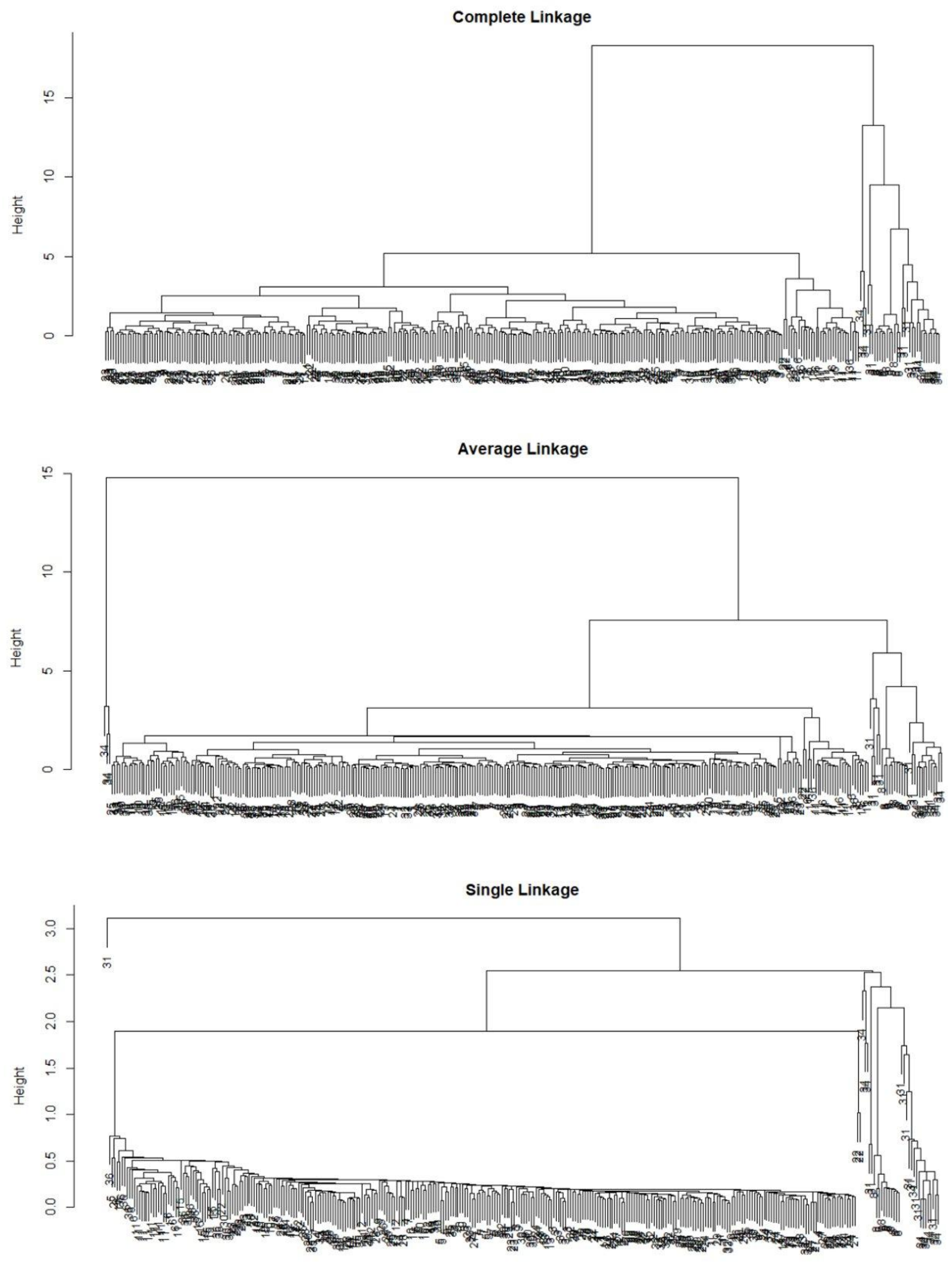

Figure 9: Results of Hierarchical Clustering using the Complete, Average, and Single methods

Conclusion

[redacted]While observing our results throughout each method, it was noted that the best method for classifying the thirty different leaves within the Leaf dataset was Linear discriminant analysis. LDA received a cross validated accuracy of 78.08% with 1,000 iterations. This accuracy was achieved using the dataset which included four interactions. Overall the interaction dataset did not perform significantly better on all modeling techniques, but did achieve the highest test accuracy on LDA.

Several alterations could be done to the technique to possibly increase overall accuracy in the future. The addition of more or different interactions could lead to better accuracy. Interactions that were included were simply picked from the chart correlations, instead picking more significant variables that were found using K-Means Cluster Analysis and Principal Component Analysis may result in better performance of modeling techniques considered. Also, the inclusion of the ten omitted leaf categories may increase the overall accuracy, simply by increasing the size of the test and training set. The additional leaf categories would also increase the range of data our model would be able to categorize, making it more practical for real world applications.