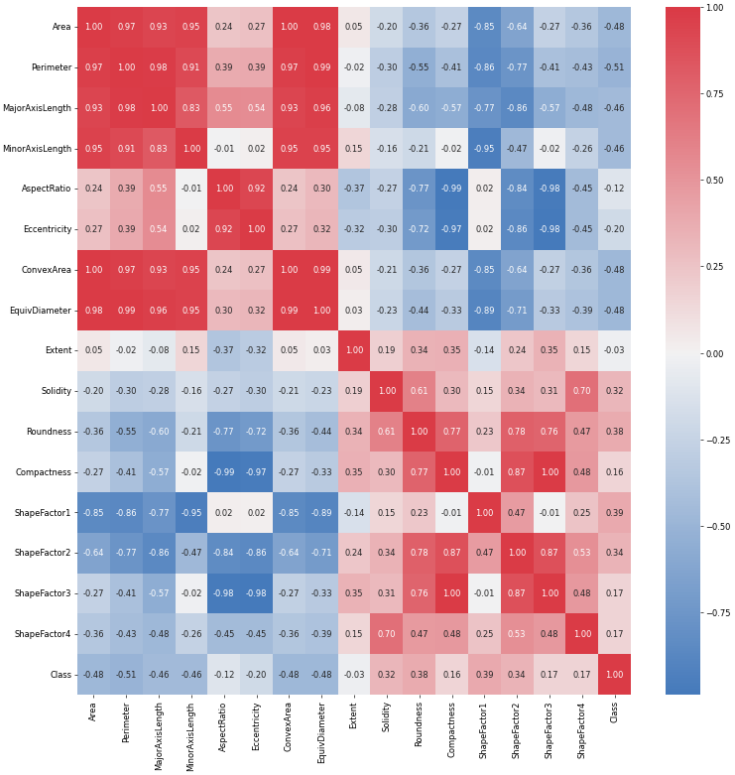
Ann Biju AXB190082
Whitney Humecky WJH190000
CS4372 HW 2

Predictors

 # predictors we included due to having highest correlations to ‘class’ (target) variable (|corr|>0.3):
Area
Perimeter
Major Axis Length
Minor Axis Length
Convex Area
EquivDiameter
Solidity
Roundness
Shape Factor 1
Shape Factor 2

Lowest. -0.3 < Correlation to class < 0.3 (exclude):
Aspect Ratio
Eccentricity
Extent
Compactness
Shape Factor 3
Shape Factor 4

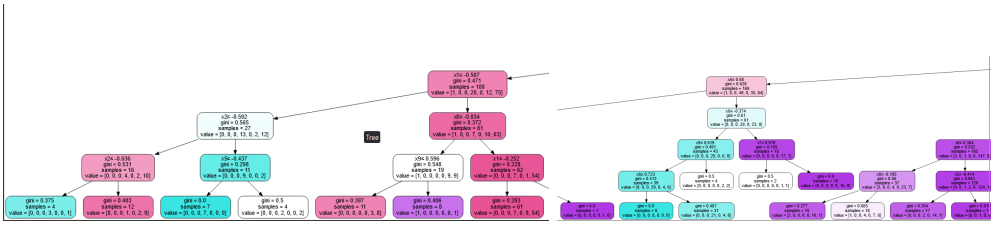


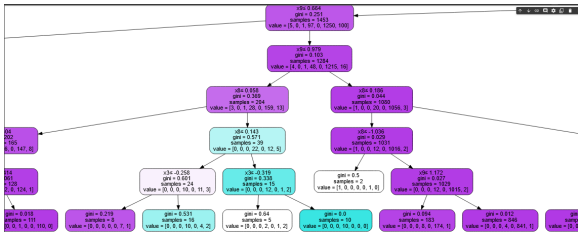
Correlation heat map showing the correlations between each predictor to every other predictor

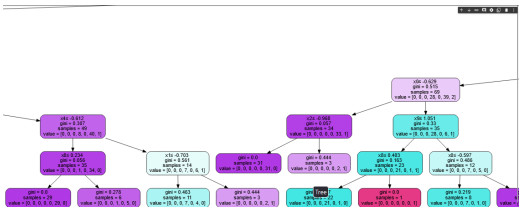
2.4 Result Analysis

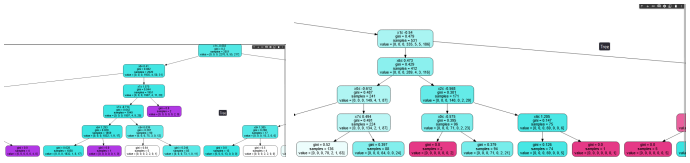
1. Plain Decision Tree Classifier

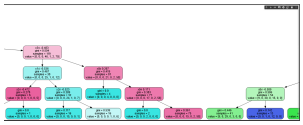
*Note: we have captured each of the trees in several snapshots, since each individual tree is too big to fit in 1 screenshot.

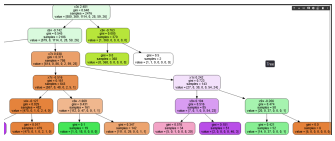


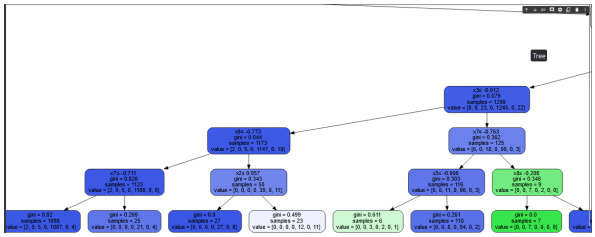


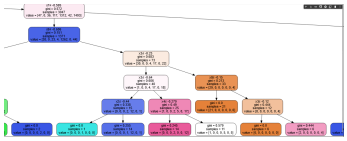


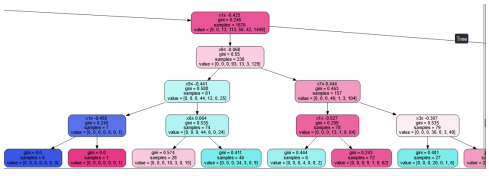


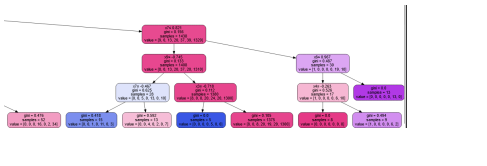












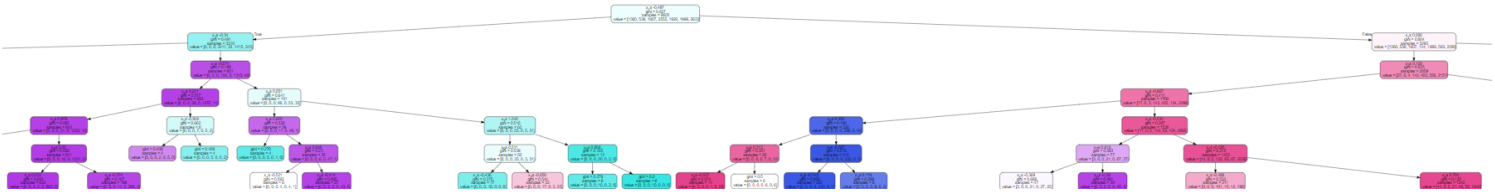
```
Confusion Matrix:
[[329   0  41   0   6   2  12]
 [  1 152   0   0   0   0   0]
 [ 36   0 422   0  14   1   6]
 [  0   0   0 999   1   9  53]
 [  7   0   5   6 545   0  16]
 [  3   0   1  19   0 575  23]
 [  4   0   0 102  15  16 683]]
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.84      0.85       390
           1       1.00      0.99      1.00       153
           2       0.90      0.88      0.89       479
           3       0.89      0.94      0.91      1062
           4       0.94      0.94      0.94       579
           5       0.95      0.93      0.94       621
           6       0.86      0.83      0.84       800

    accuracy                           0.90      4084
   macro avg       0.91      0.91      0.91      4084
weighted avg       0.90      0.90      0.90      4084
```

The decision tree is a strong model, not the strongest of the four produced here but still impressive at weighted averages of 90% for precision, recall, and F1 score - meaning regardless of the predicted class (0-6), the model is effectively accurate. Precision for each class ranges from 86% to 100% indicating that when the model assigns positive, it is 90% accurate. The recall range is slightly weaker at 83% - 99% meaning that the model rarely misassigns a negative class. The confusion matrix has a strong diagonal of true positives and negatives - again, there are stronger models created below so this is not the model of choice today for the Bean dataset.

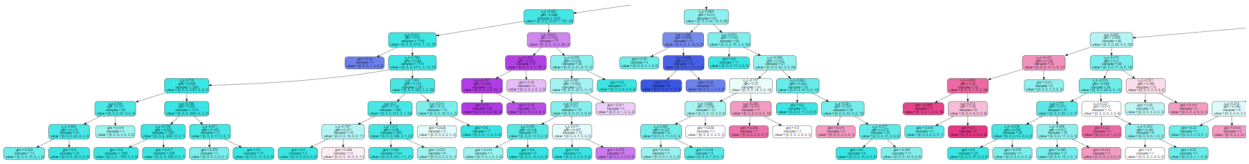
2. Random Forest Tree Classifier

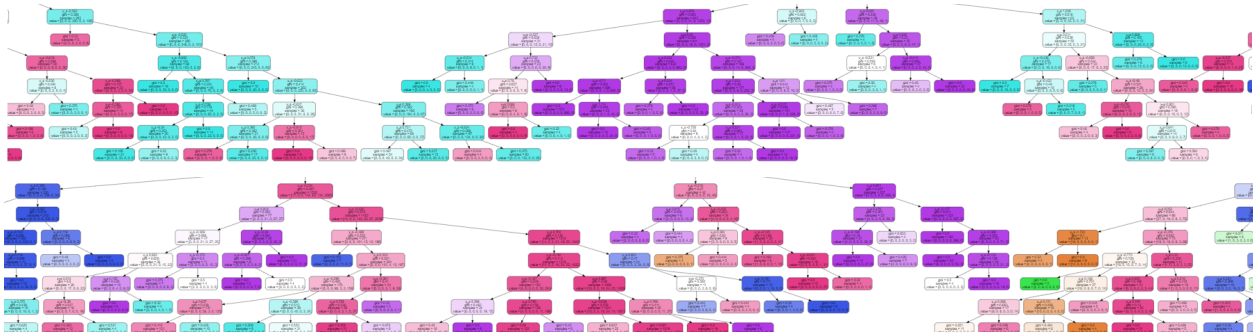
Tree:

Layer 1:



Layer 2:



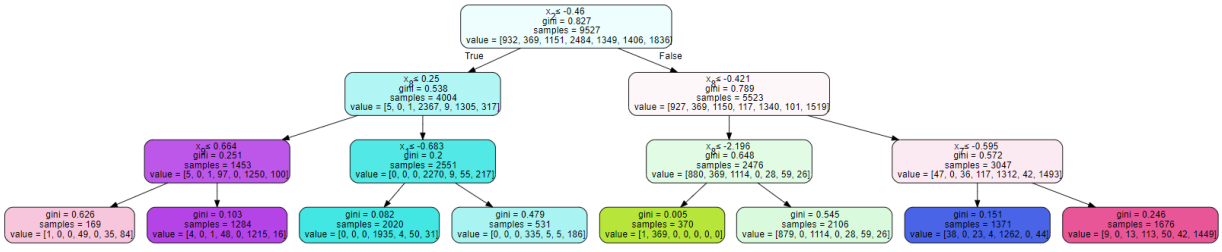


```
Confusion Matrix:
[[ 341    0   34    0    2    2   11]
 [   0  153    0    0    0    0    0]
 [   9    0  462    0    6    1    1]
 [   0    0    0 1012    0    7   43]
 [   3    0    5    4  556    0   11]
 [   3    0    0   12    0  594   12]
 [   1    0    0   70    6    8  715]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.87      0.91       390
           1       1.00      1.00      1.00       153
           2       0.92      0.96      0.94       479
           3       0.92      0.95      0.94      1062
           4       0.98      0.96      0.97       579
           5       0.97      0.96      0.96       621
           6       0.90      0.89      0.90       800

    accuracy                           0.94      4084
   macro avg       0.95      0.94      0.95      4084
weighted avg       0.94      0.94      0.94      4084
```

Analysis of Results:
Accuracy of the Random Forest model comes in at 94%, tying with the XGBoost Classifier model. Precision ranges from 0.90 to 1.00, indicating a very precise model. While the confusion matrix shows strong diagonals, there are some sparse values within. This model comes in at 2nd best.

3. Adaboost Classifier:

Tree:

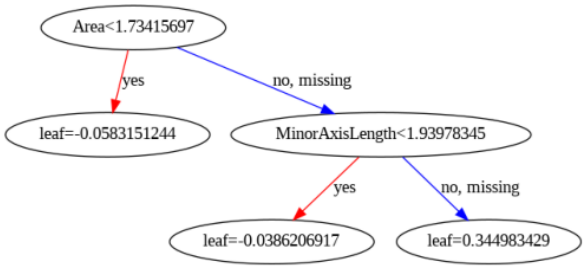


```
Confusion Matrix:
[[  0   0 361   0  22   0   7]
 [  0 153   0   0   0   0   0]
 [  0   0 463   0   9   0   7]
 [  0   0   0 985   1  11  65]
 [  0   0   8   5 541   0  25]
 [  0   0  39  33   0 516  33]
 [  0   0   9  97  21   9 664]]
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       390
           1       1.00      1.00      1.00       153
           2       0.53      0.97      0.68       479
           3       0.88      0.93      0.90      1062
           4       0.91      0.93      0.92       579
           5       0.96      0.83      0.89       621
           6       0.83      0.83      0.83       800

    accuracy                           0.81      4084
   macro avg       0.73      0.78      0.75      4084
weighted avg       0.77      0.81      0.78      4084
```

The Adaboost is the weakest model created for the Bean dataset. While its weighted average precision, recall, and f1 -score are still adequate at above 75%, it could be better. Only 77% of the time, the assigned positives are true. 81% of the time the true positive values are identified as such, and thus the balance between precision and recall is only 81%. Precision has a wide range of 53% to 100% - this model is particularly inaccurate when handling a node of class 2.

4. XGBoost Classifier

Tree:



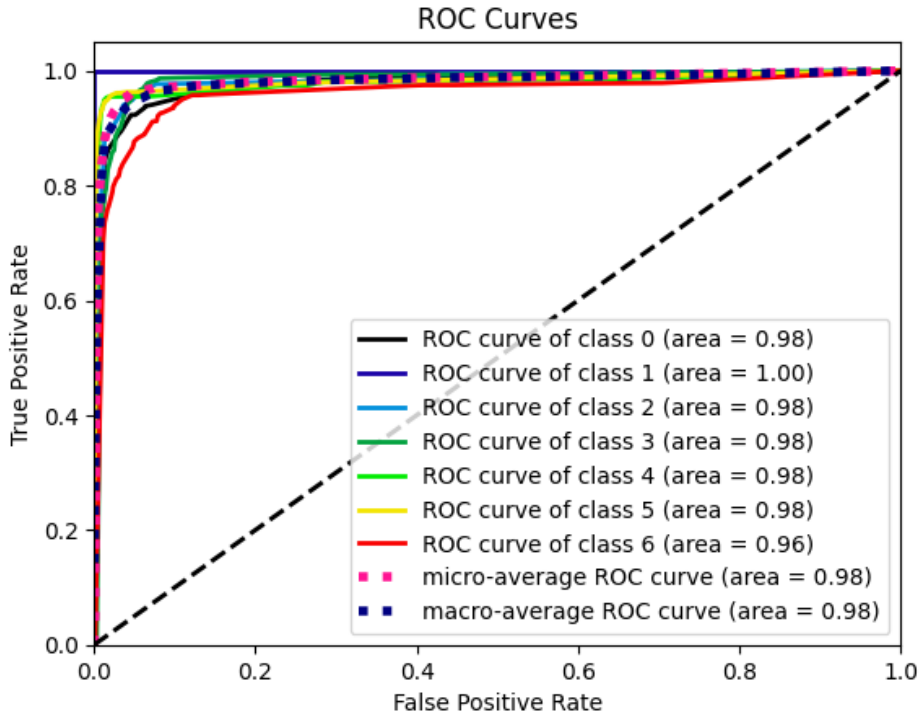

```
Confusion Matrix:
[[ 355    0   25    0    1    2    7]
 [   0  153    0    0    0    0    0]
 [   6    1  465    0    6    0    1]
 [   0    0    0 1004    0    8   50]
 [   3    0    2    3  560    0   11]
 [   1    0    0    8    0  598   14]
 [   0    0    0   71    7   11  711]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.91      0.94       390
           1       0.99      1.00      1.00       153
           2       0.95      0.97      0.96       479
           3       0.92      0.95      0.93      1062
           4       0.98      0.97      0.97       579
           5       0.97      0.96      0.96       621
           6       0.90      0.89      0.89       800

    accuracy                           0.94      4084
   macro avg       0.95      0.95      0.95      4084
weighted avg       0.94      0.94      0.94      4084
```

The weighted average precision, recall, and F1 score for the XGBoost model are all extremely high at 94%. This means the model is incredibly strong when predicting positive classes, rarely assigning false positives or negatives, and that there is a good balance between precision and recall. The confusion matrix has a heavy diagonal where true positives and true negatives are reported, and sparse upper and lower triangles - where false positives and negatives are reported - again indicating a strong model.

Best Model - XGBoost

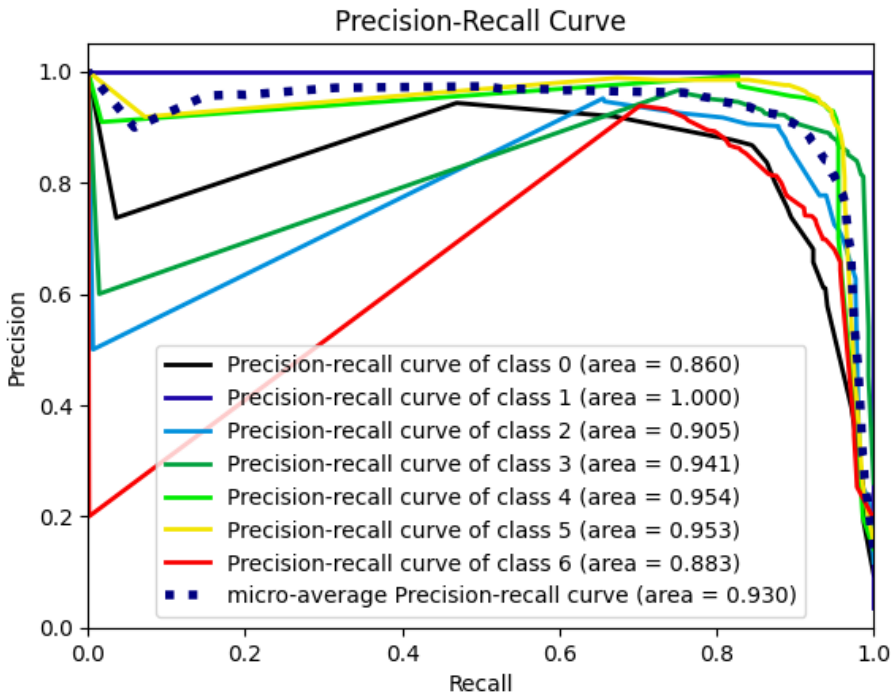
The XGBoost and Random Forest models had incredibly comparable measurements. XGBoost is a much simpler model, making it much easier to work with, faster to process, and our best model of the four. The XGBoost model had narrow ranges of less than 11% for precision, recall, and f1-scores making it a good model for any of the 7 classes.

ROC:



Class 1 has the highest ROC AUC (1), through which diagnostic we can safely say that this class has a higher probability that a randomly chosen positive instance is ranked higher than that of a negative instance.

Precision-Recall Curve:



Class 0 seems to show the smallest tradeoff between precision and recall, making it a strong classifier, having the highest precision (1.0), while also maintaining a recall (or sensitivity rating) close to 1.0. Class 0 is the only class that maintains its curve above the baseline average Precision-Recall indicated by the dotted line.