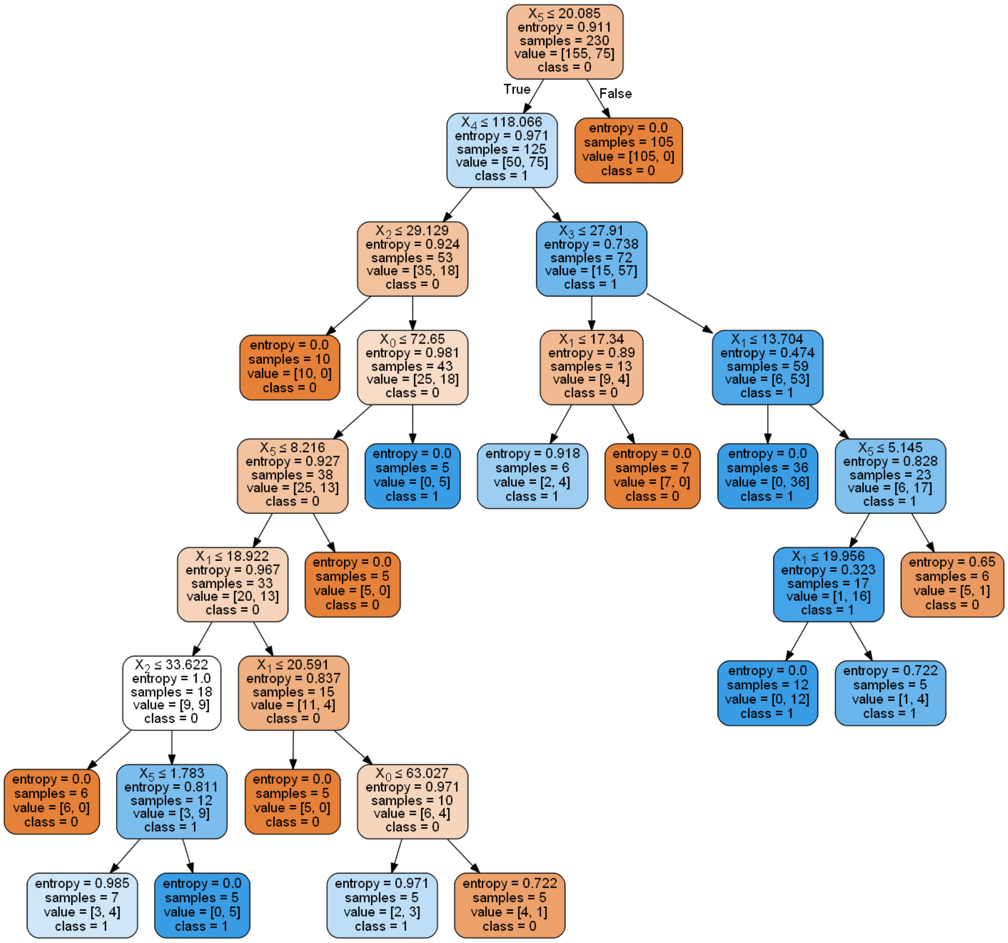
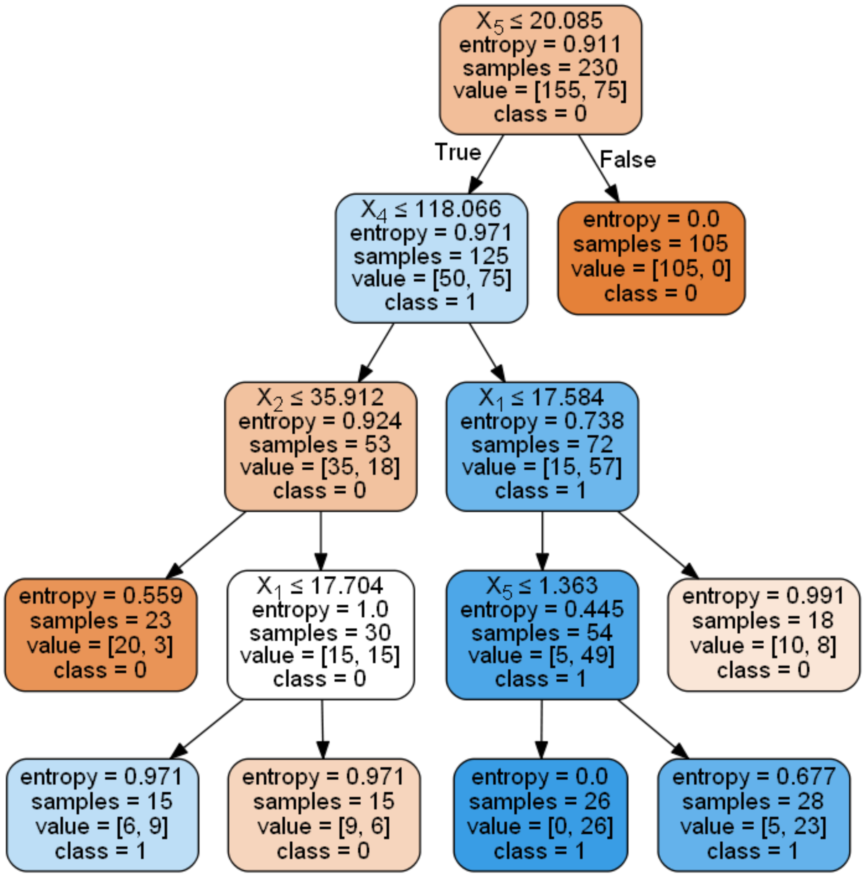
Decision Tree Analysis Based on Entropy Values

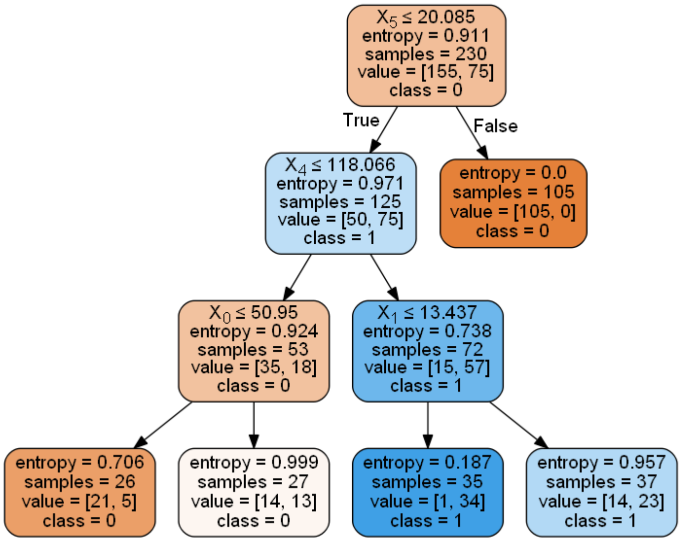
Minimum 5 sample leaf nodes



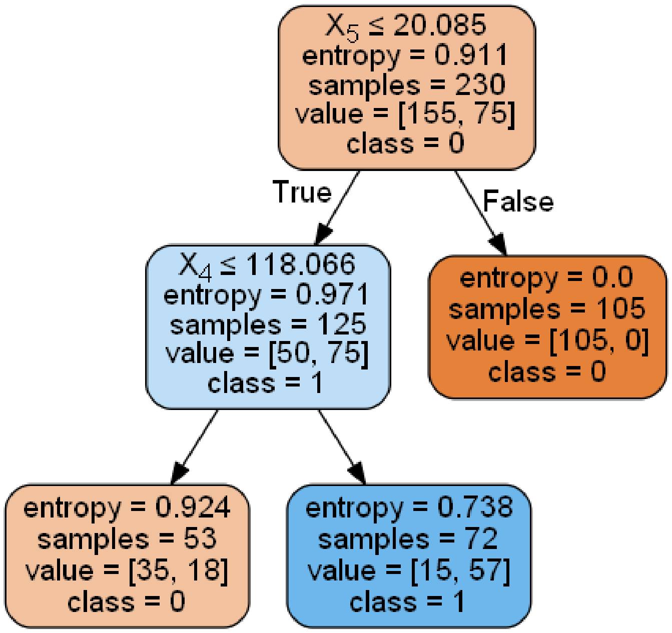
15 samples leaf node



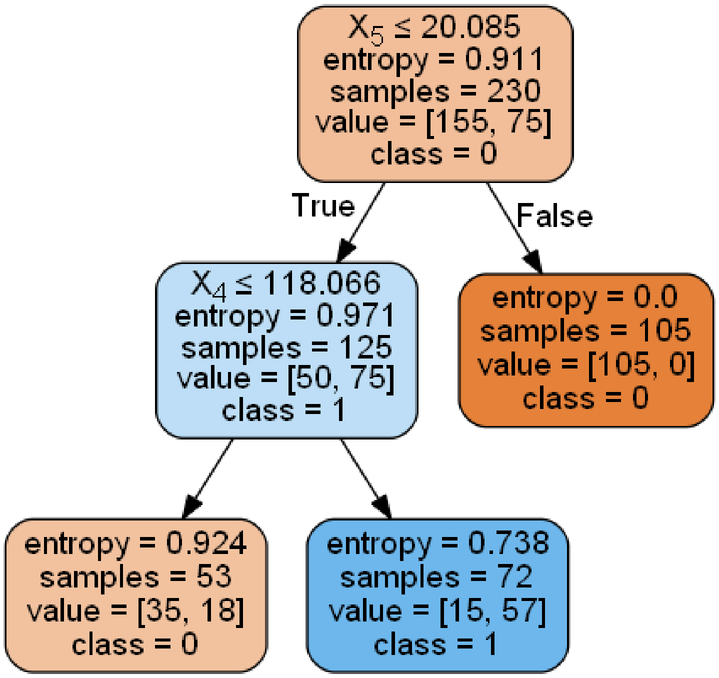
25 sample leaf node



Min 40 sample leaf node:



Min 50 sample leaf node



Conclusion:

Among the 5 decision trees, I would pick the decision tree with the min 15 samples in the leaf node. As we increase min sample value in leaf node the decision tree shallower, However, if we check the impurity value entropy values are higher in the shallow trees. This shows that we can't achieve our goal maximizing the information gain. Although the tree with min 5 sample in leaf node gives the good information gain, It is not very shallow and simple. I don't see huge information gain difference between 5 sample leaf and 15 sample leaf. As considering Occam's Razor, simpler model rule. I chose the tree with min 15 samples in its leaf nodes.

Accuracy and Precision Values

Minimum 5 sample leaf nodes

```
          precision    recall  f1-score   support

        0      0.88      0.82      0.85        55
        1      0.66      0.76      0.70        25

  accuracy                         0.80        80
 macro avg       0.77      0.79      0.78        80
weighted avg     0.81      0.80      0.80        80
```

Minimum 15 sample leaf nodes

```
          precision    recall  f1-score   support

        0      0.87      0.84      0.85        55
        1      0.67      0.72      0.69        25

  accuracy                         0.80        80
 macro avg       0.77      0.78      0.77        80
weighted avg     0.81      0.80      0.80        80
```

Minimum 25 sample leaf nodes

```
          precision    recall  f1-score   support

        0      0.86      0.78      0.82        55
        1      0.60      0.72      0.65        25

  accuracy                         0.76        80
 macro avg       0.73      0.75      0.74        80
weighted avg     0.78      0.76      0.77        80
```

Minimum 40 sample leaf nodes

```
              precision    recall  f1-score   support

           0       0.86      0.78      0.82        55
           1       0.60      0.72      0.65        25

    accuracy                           0.76        80
   macro avg       0.73      0.75      0.74        80
weighted avg       0.78      0.76      0.77        80
```

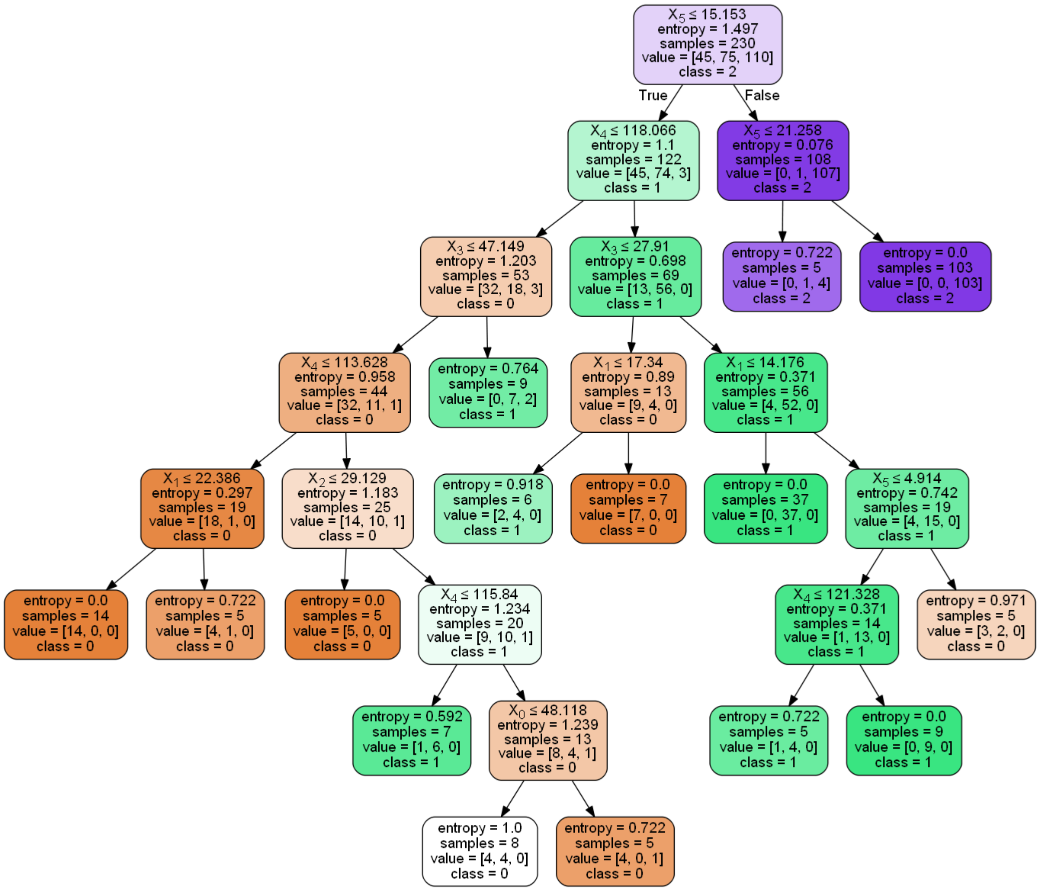
Minimum 50 sample leaf nodes

```
              precision    recall  f1-score   support

           0       0.86      0.78      0.82        55
           1       0.60      0.72      0.65        25

    accuracy                           0.76        80
   macro avg       0.73      0.75      0.74        80
weighted avg       0.78      0.76      0.77        80
```

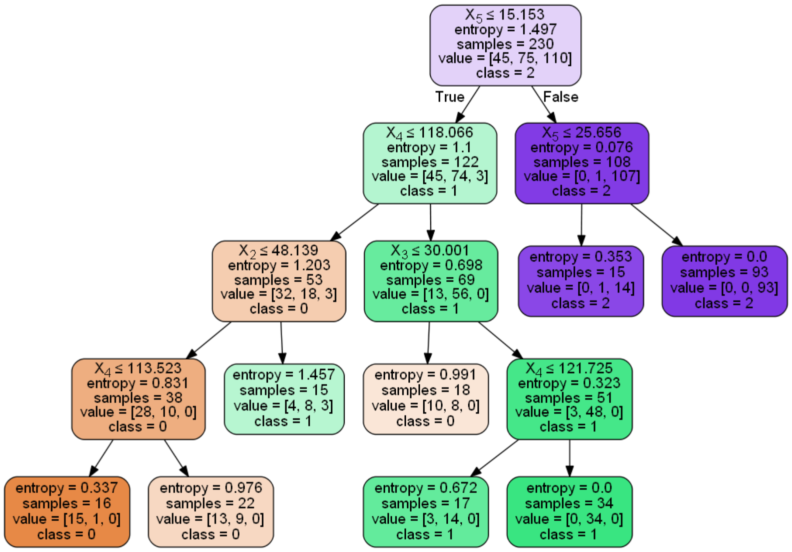
5 and 15 sample leaf node models have the same accuracy rate 80% while others show the 76 % of accuracy. Moreover, 5 and 15 sample leaf node models are higher precision and recall values that give us better exactness and relevancy.

Repeating Analysis with Having Three Different Class Labels

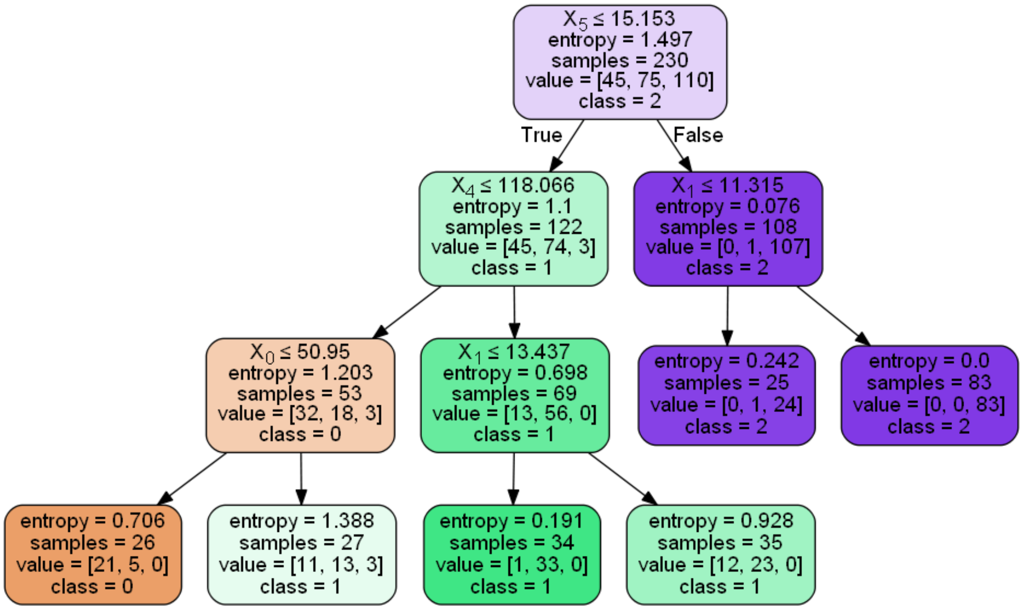
Min 5 sample leaf node



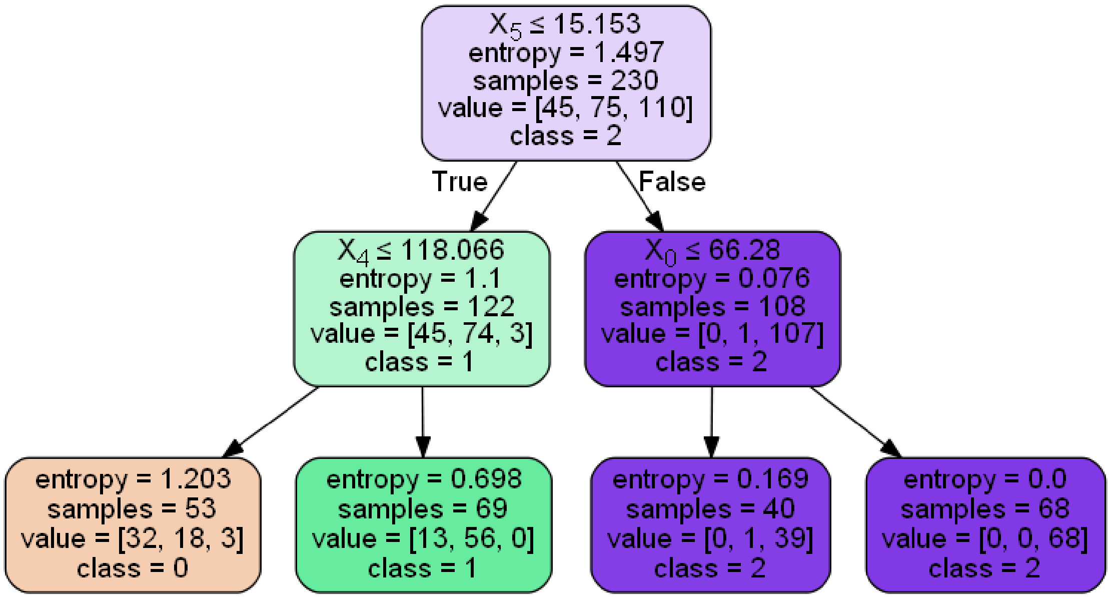
Min 15 samples leaf node



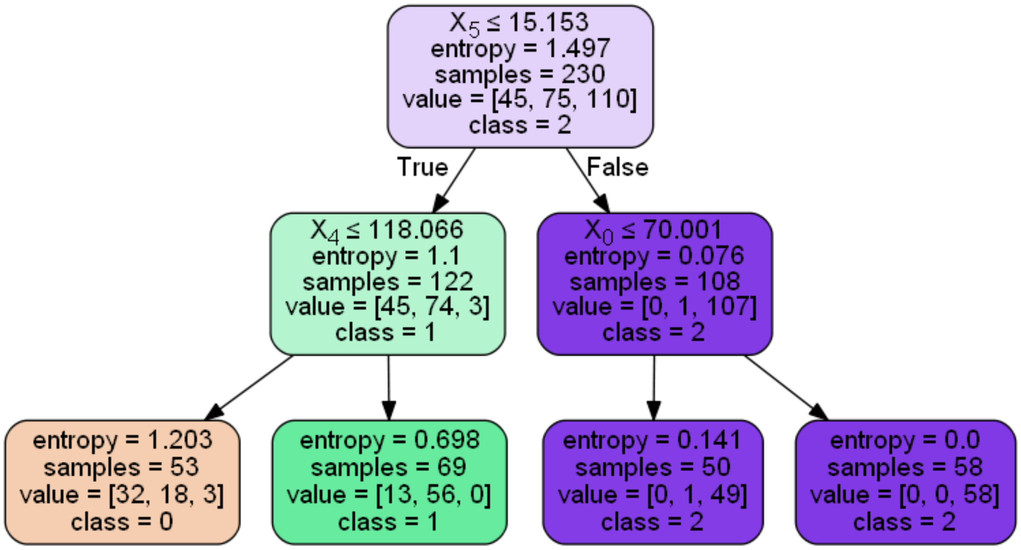
Min 25 samples leaf node



Min 40 samples leaf node



Min 50 samples leaf node



Conclusion:

Min 5 sample decision tree is shallower in this decision tree. I think, adding one more class helps for classifying better in this data set. On the other hand, in each every decision trees' leaf node, entrophy is very high for the class 0. So we don't have information gain for class 0. In every decision tree, class 2 is easily splitted. Also, compared to the decision tree in the previous example, I would still select the decision tree with min 15 sample leaf node because of the precision recall accuracy and weighted average values. I was leaning towards to select 25 samples one however its weighted average value and accuracy is very low.

Accuracy and Precision Values for 3 different class labels scenario

Min 5 samples leaf node

```
          precision   recall  f1-score   support

       0      0.50      0.47      0.48        15
       1      0.68      0.68      0.68        25
       2      0.93      0.95      0.94        40

   accuracy                       0.78        80
  macro avg      0.70      0.70      0.70        80
weighted avg      0.77      0.78      0.77        80
```

Min 15 samples leaf node

```
          precision   recall  f1-score   support

       0      0.59      0.67      0.62        15
```

```
           1       0.73      0.64      0.68        25
           2       0.93      0.95      0.94        40

    accuracy                           0.80        80
   macro avg       0.75      0.75      0.75        80
weighted avg       0.80      0.80      0.80        80
```

Min 25 samples leaf node

```
              precision    recall  f1-score   support

           0       0.08      0.27      0.12        15
           1       0.60      0.72      0.65        25
           2       0.00      0.00      0.00        40

    accuracy                           0.28        80
   macro avg       0.23      0.33      0.26        80
weighted avg       0.20      0.28      0.23        80
```

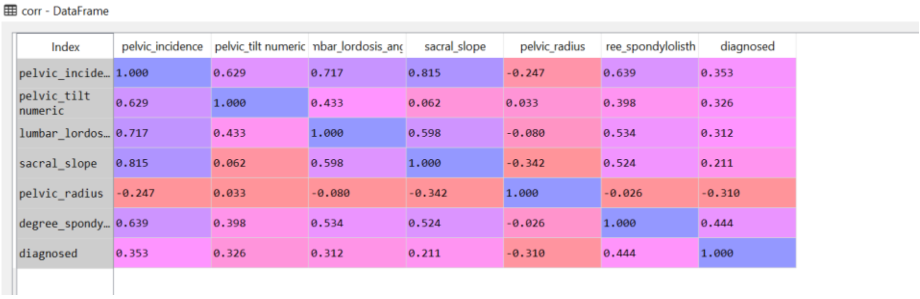
Min 40 samples leaf node

```
              precision    recall  f1-score   support

           0       0.33      0.20      0.25        15
           1       0.60      0.72      0.65        25
           2       0.93      0.95      0.94        40

    accuracy                           0.74        80
   macro avg       0.62      0.62      0.61        80
weighted avg       0.71      0.74      0.72        80
```

Min 50 samples leaf node

```
              precision    recall  f1-score   support

           0       0.33      0.20      0.25        15
           1       0.60      0.72      0.65        25
           2       0.93      0.95      0.94        40

    accuracy                           0.74        80
   macro avg       0.62      0.62      0.61        80
weighted avg       0.71      0.74      0.72        80
```

Eliminating one of the variable that has high correlation with another independent variable and repeating decision tree analysis

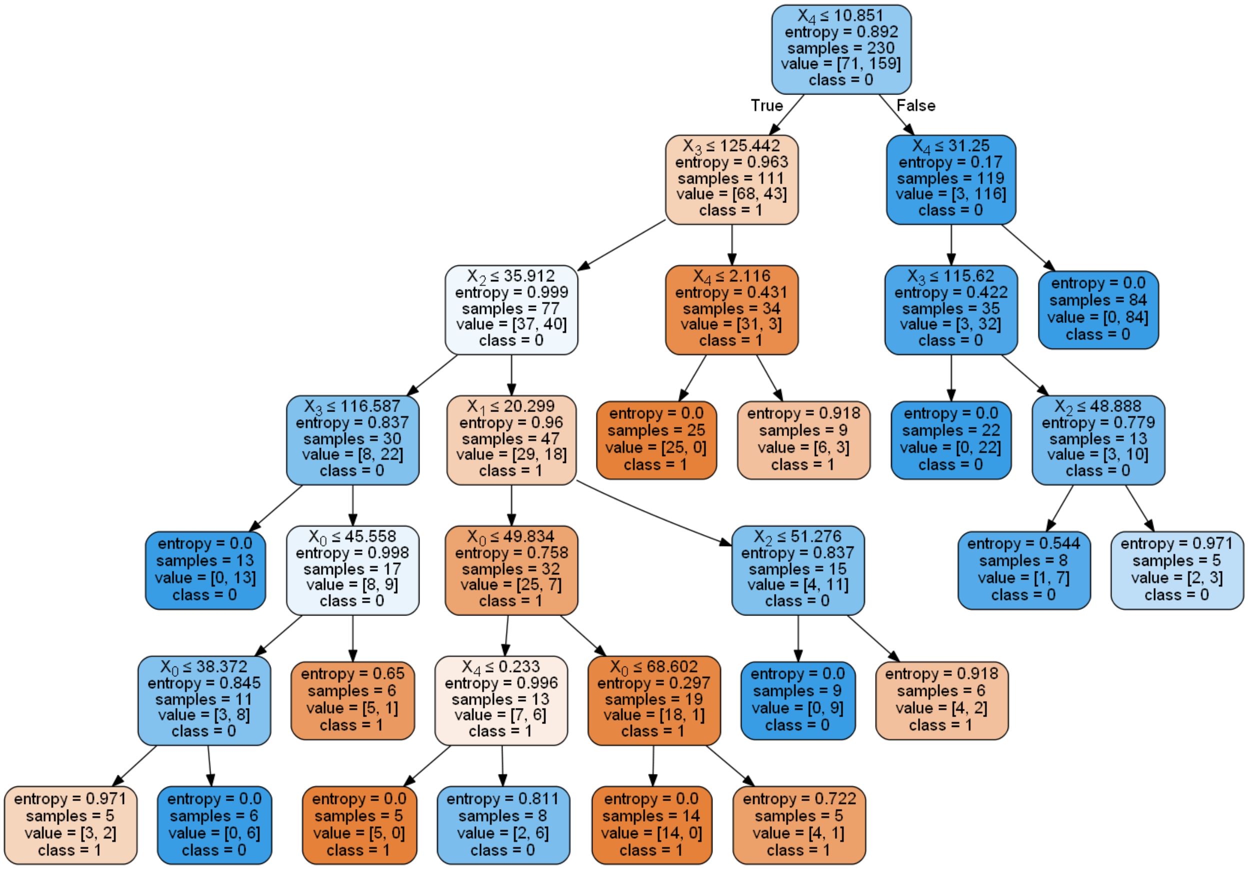


corr - DataFrame

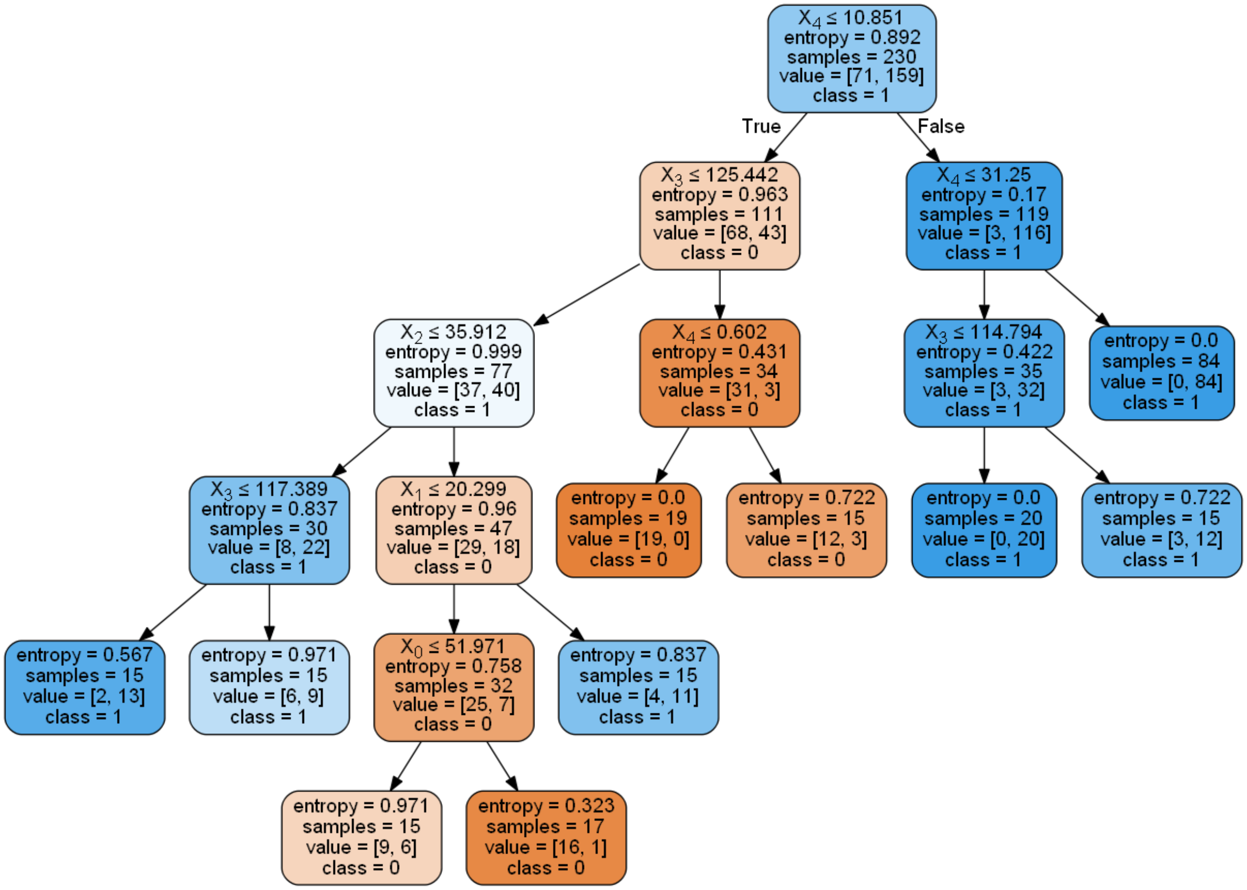
Index	pelvic_incidence	pelvic_tilt numeric	nbar_lordosis_anç	sacral_slope	pelvic_radius	ree_spondylolisth	diagnosed
pelvic_incide…	1.000	0.629	0.717	0.815	-0.247	0.639	0.353
pelvic_tilt numeric	0.629	1.000	0.433	0.062	0.033	0.398	0.326
lumbar_lordos…	0.717	0.433	1.000	0.598	-0.080	0.534	0.312
sacral_slope	0.815	0.062	0.598	1.000	-0.342	0.524	0.211
pelvic_radius	-0.247	0.033	-0.080	-0.342	1.000	-0.026	-0.310
degree_spondy…	0.639	0.398	0.534	0.524	-0.026	1.000	0.444
diagnosed	0.353	0.326	0.312	0.211	-0.310	0.444	1.000

I eliminated the variable sacral_slope since it has high correaltion with pelvis_incidence.

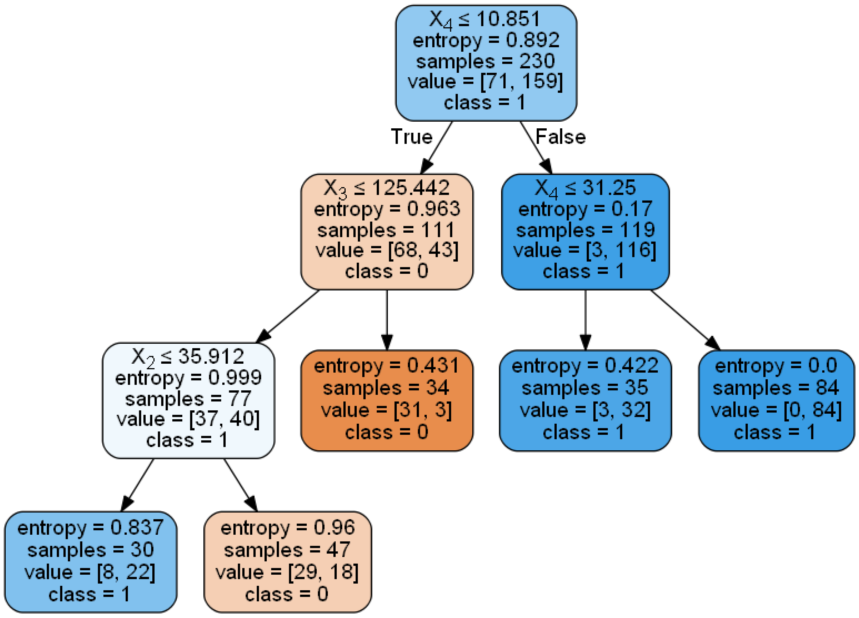
Min 5 samples leaf node


$X_4 \leq 10.851$ entropy = 0.892 samples = 230 value = [71, 159] class = 0
True
False
$X_3 \leq 125.442$ entropy = 0.963 samples = 111 value = [68, 43] class = 1
$X_4 \leq 31.25$ entropy = 0.17 samples = 119 value = [3, 116] class = 0
$X_2 \leq 35.912$ entropy = 0.999 samples = 77 value = [37, 40] class = 0
$X_4 \leq 2.116$ entropy = 0.431 samples = 34 value = [31, 3] class = 1
$X_3 \leq 115.62$ entropy = 0.422 samples = 35 value = [3, 32] class = 0
entropy = 0.0 samples = 84 value = [0, 84] class = 0
$X_3 \leq 116.587$ entropy = 0.837 samples = 30 value = [8, 22] class = 0
$X_1 \leq 20.299$ entropy = 0.96 samples = 47 value = [29, 18] class = 1
entropy = 0.0 samples = 25 value = [25, 0] class = 1
entropy = 0.918 samples = 9 value = [6, 3] class = 1
entropy = 0.0 samples = 22 value = [0, 22] class = 0
$X_2 \leq 48.888$ entropy = 0.779 samples = 13 value = [3, 10] class = 0
entropy = 0.0 samples = 13 value = [0, 13] class = 0
$X_0 \leq 45.558$ entropy = 0.998 samples = 17 value = [8, 9] class = 0
$X_0 \leq 49.834$ entropy = 0.758 samples = 32 value = [25, 7] class = 1
$X_2 \leq 51.276$ entropy = 0.837 samples = 15 value = [4, 11] class = 0
entropy = 0.544 samples = 8 value = [1, 7] class = 0
entropy = 0.971 samples = 5 value = [2, 3] class = 0
$X_0 \leq 38.372$ entropy = 0.845 samples = 11 value = [3, 8] class = 0
entropy = 0.65 samples = 6 value = [5, 1] class = 1
$X_4 \leq 0.233$ entropy = 0.996 samples = 13 value = [7, 6] class = 1
$X_0 \leq 68.602$ entropy = 0.297 samples = 19 value = [18, 1] class = 1
entropy = 0.0 samples = 9 value = [0, 9] class = 0
entropy = 0.918 samples = 6 value = [4, 2] class = 1
entropy = 0.971 samples = 5 value = [3, 2] class = 1
entropy = 0.0 samples = 6 value = [0, 6] class = 0
entropy = 0.0 samples = 5 value = [5, 0] class = 1
entropy = 0.811 samples = 8 value = [2, 6] class = 0
entropy = 0.0 samples = 14 value = [14, 0] class = 1
entropy = 0.722 samples = 5 value = [4, 1] class = 1

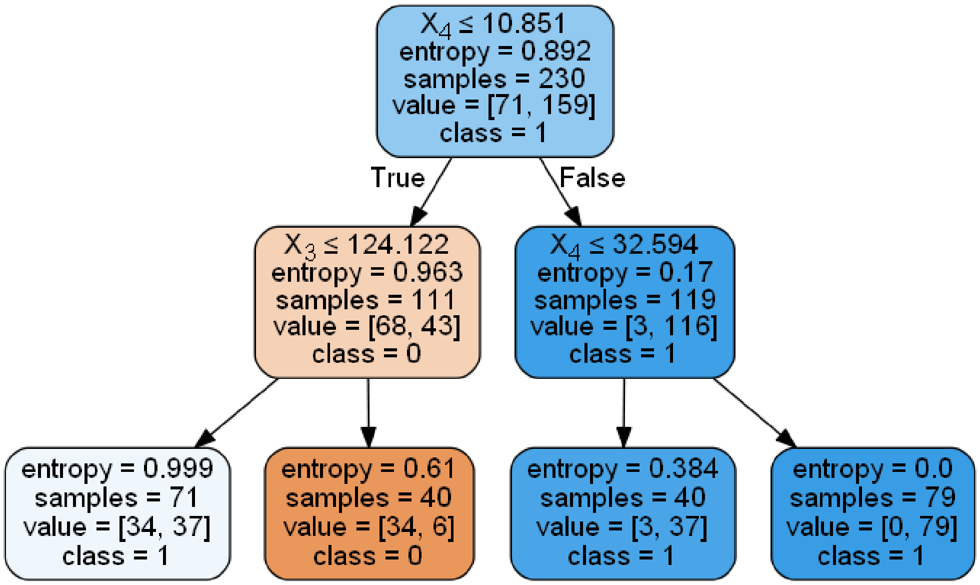
Min 15 samples leaf node



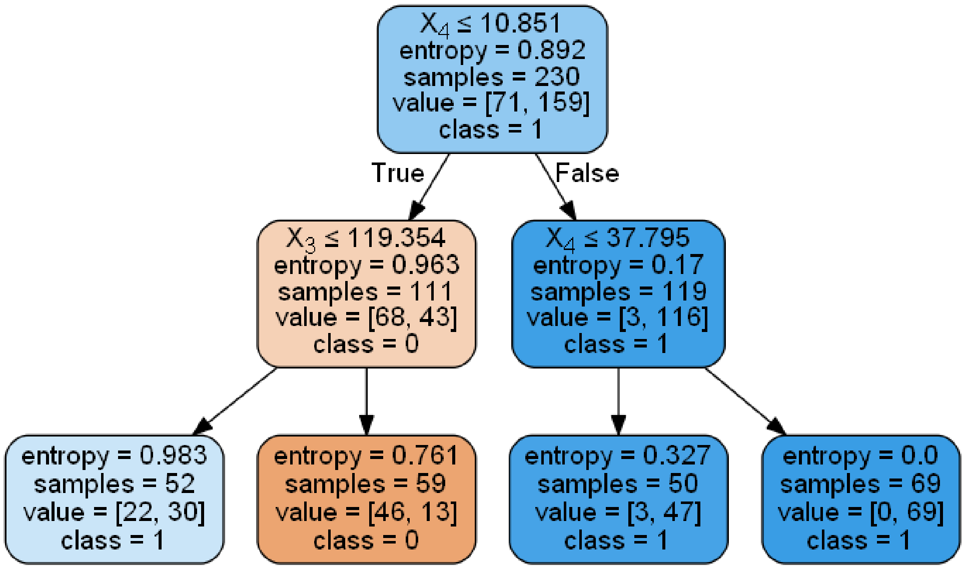
Min 25 samples leaf node



Min 40 samples leaf node



Min 50 samples leaf node



Accuracy and Precision Values

Min 5 samples leaf node

0	0.74	0.69	0.71	29
1	0.83	0.86	0.85	51

```
   accuracy                         0.80        80
  macro avg       0.79      0.78      0.78        80
weighted avg       0.80      0.80      0.80        80
```

Min 15 samples leaf node

```
              precision    recall  f1-score   support

           0       0.75      0.62      0.68        29
           1       0.80      0.88      0.84        51

   accuracy                         0.79        80
  macro avg       0.78      0.75      0.76        80
weighted avg       0.78      0.79      0.78        80
```

Min 25 samples leaf node

```
              precision    recall  f1-score   support

           0       0.71      0.69      0.70        29
           1       0.83      0.84      0.83        51

   accuracy                         0.79        80
  macro avg       0.77      0.77      0.77        80
weighted avg       0.79      0.79      0.79        80
```

Min 50 samples leaf node

```
              precision    recall  f1-score   support

           0       0.73      0.38      0.50        29
           1       0.72      0.92      0.81        51

   accuracy                         0.73        80
  macro avg       0.73      0.65      0.66        80
weighted avg       0.73      0.72      0.70        80

              precision    recall  f1-score   support

           0       0.69      0.62      0.65        29
           1       0.80      0.84      0.82        51

   accuracy                         0.76        80
  macro avg       0.74      0.73      0.74        80
weighted avg       0.76      0.76      0.76        80
```

Conclusion :

When I left the variable sacral_slope, accuracy increased for decision tree with 25 min sample leaf and above others. Also, the decision tree depth for 5 min sample became little shallower. Comparatively, entropy values are in leaves are smaller than before we eliminated the correlated value. Since our sample size is not large, the change in the accuracy is not very big but we can still observe it.