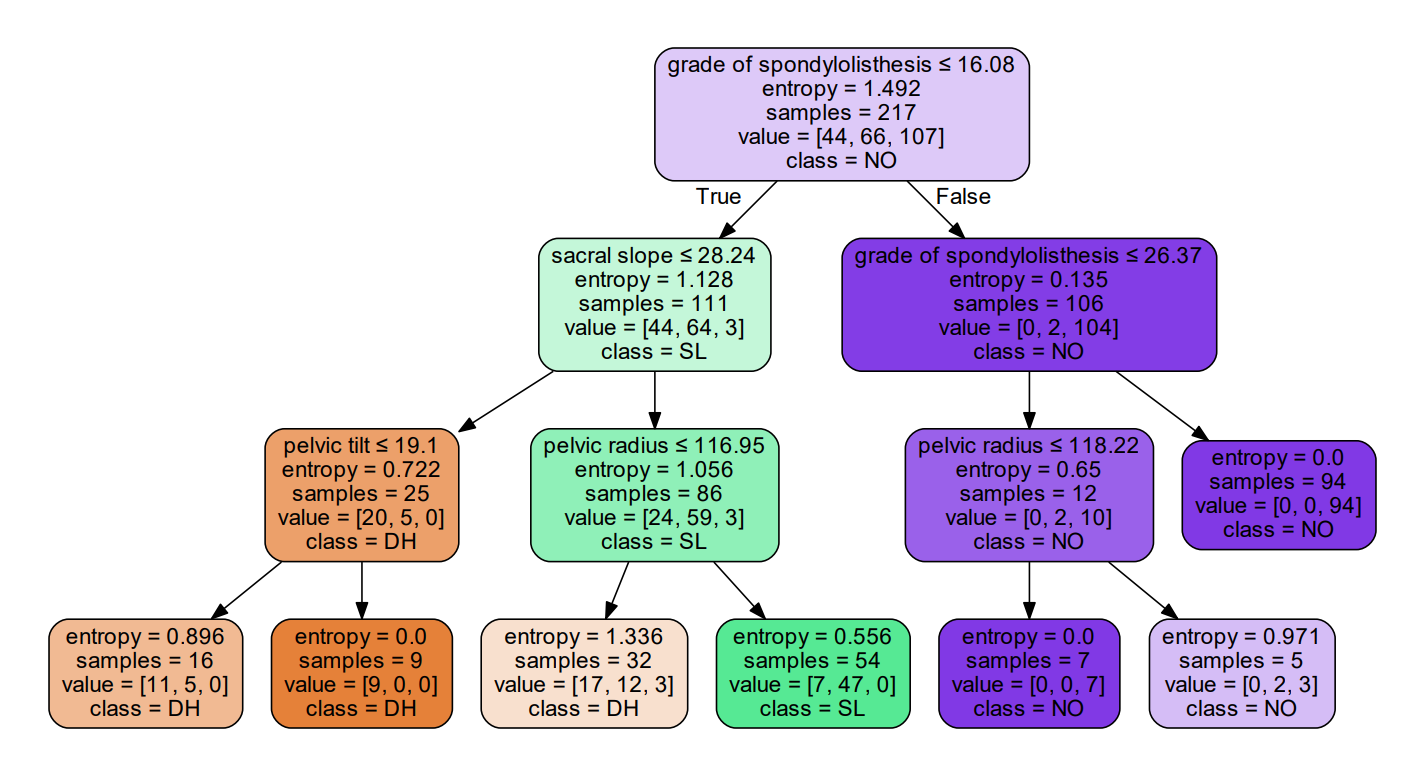
**Assignment Report**

1. Three trees will be built up. The dataset is split in a ratio of 70:30 between training and testing. 70% of the dataset will be used for training to build up and decision tree and then 30% of the rest dataset which should be enough will be used for validation to evaluate the decision tree’s accuracy. Therefore, we pass a parameter 0.3 as test\_size. Only entropy will be used as impurity measure to test the dataset because we found that it provides a slightly higher accuracy than the Gini index. Random\_state is set to be 100 to ensure that every time the same dataset is tested. The only difference is that the maximum depth for these three trees will be 3, 4 and 5.
2. For the first tree, by using entropy, the accuracy is about 80%. It contains a depth of 3. For the second tree, the classification performance using entropy is about 85%. It contains a depth of 4. For the third tree, the classification performance using entropy is about 83%. It contains a depth of 5. Therefore, the third tree with a depth of 4 has the highest accuracy. When it comes to a depth of 5, overfitting occurs, and the accuracy drops. However, when the depth is 3, underfitting occurs, the accuracy is not the maximum. Therefore, when the tree with a depth of 3, it is the most suitable depth and maximize the accuracy.
3. For the first selected tree, by using entropy, the precision of the class DH is 0.5. The precision of the class SL and NO are 0.81 and 0.98 respectively. The NO class has the highest precision. From the confusion matrix, we can see that 4 of the samples belonging to class DH have been misclassified as class SL. 11 of the samples belonging to class SL are misclassified as class DH and 1 is misclassified as class NO. While for the class NO, only 1 is misclassified as class DH and 1 as class SL. The samples are mainly misclassified as class DH or SL. Therefore, the pairs of class SL and DH are likely to be confused with each other.

For the second selected tree, shown from the classification report by using entropy, the precision of class DH, SL and NO are 0.63, 0.81 and 0.98 respectively. The NO class has the highest precision. From the confusion matrix, we can see that 4 of the samples belonging to the class DH have been misclassified as class SL. 7 of the samples belonging to class SL are misclassified as class DH and 1 is misclassified as class NO. While for the class NO, only 2 is misclassified as class SL. The samples are mainly misclassified as class DH or SL. Therefore, the pairs of class SL and DH are likely to be confused with each other.

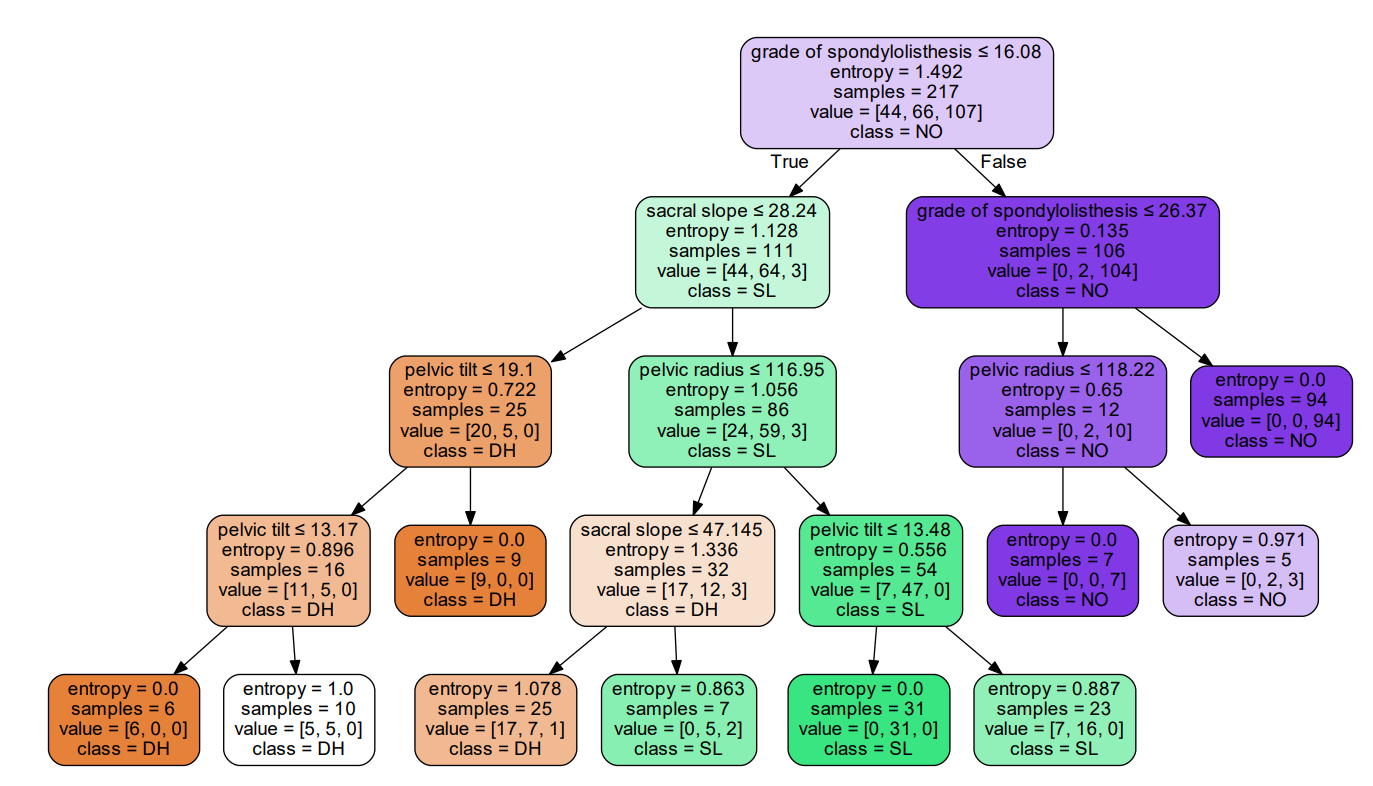
For the third selected tree, shown from the classification report by using entropy, the precision of class DH, SL and NO are 0.60, 0.75 and 0.98 respectively. The NO class has the highest precision. From the confusion matrix, we can see that 7 of the samples belonging to the class DH are misclassified as class SL. 6 of the samples belonging to the SL class are misclassified as class DH. 1 of the samples belonging to the SL class is misclassified as class NO. And 2 of the samples belonging to the class NO are misclassified as class SL. The samples are mainly misclassified as class DH or SL. Therefore, the pairs of class SL and DH are likely to be confused with each other.

1. The plot for the first selected tree using entropy:



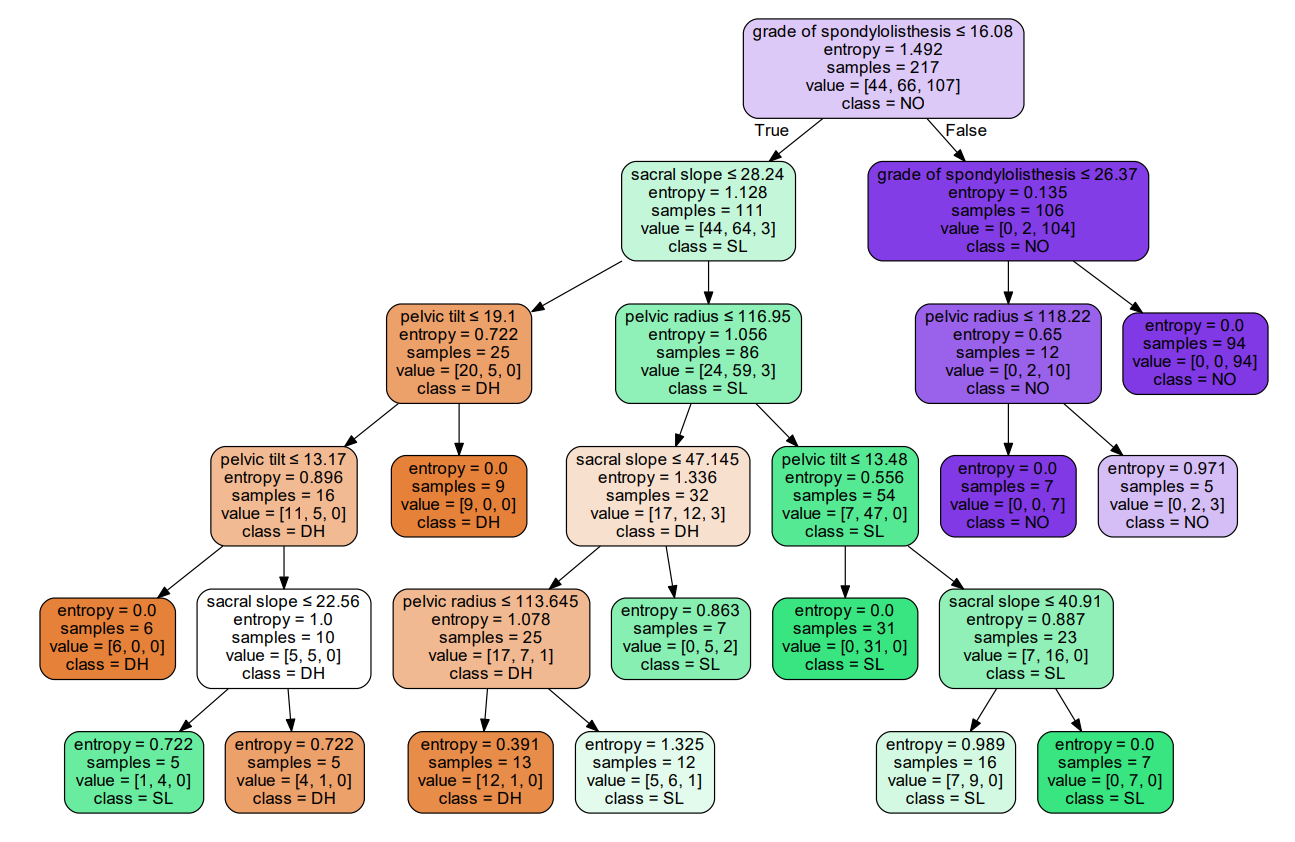
From the plot, we can see that it follows the same sequence of decisions that the grad of spondylolisthesis is less than or equal to 16.08. The difference that leads to the misclassification is the value of the sacral slope, pelvic tilt and pelvic radius. If the sacral slope is less than and equal to 28.24 and the pelvic tilt is less than or equal to 19.1, 5 samples belonging to the class SL are misclassified as class DH. Moreover, when the sacral slope is larger than 28.24 the pelvic radius is less than or equal 116.95, 12 samples belonging to the class SL are misclassified as class DH. When it is larger than 116.95, 7 samples belonging to the class DH are misclassified as class SL. From the plot, we can see that when it comes to the depth 2, if the pelvic radius is less than or equal to 116.95 or not, it cannot separate clearly the class is as SL or DH.

The plot for the second selected tree using entropy:



Since the depth of the tree is 4 which is deeper than the above tree, it follows the same sequence of decisions but provides one more row of decision making. From those leaf nodes, we can see that if the pelvic tilt is larger than 13.17, 5 samples belonging to the class SL are misclassified as class DH. Moreover, when the sacral slope is less than or equal to 47.145, 7 samples of the class SL are misclassified as class DH. Last but not least, the leaf node that the pelvic tilt > 13.48, 7 samples of the class DH are misclassified as class SL. From the plot, we can see that when it comes to the depth 3, the two nodes that if the sacral slope is less than or equal to 47.145 or not and the pelvic tilt <= 13.48 or not, it cannot separate clearly the class is as SL or DH.

The plot for the third selected tree using entropy:



Since the depth of the tree is 5 which is deeper than the above trees, it follows the same sequence of decisions but provides one more row of decision making. From those leaf nodes, we can see that if the sacral slope is <= 22.56 or not, 5 samples belonging to the class SL are misclassified as class DH. Moreover, when the pelvic radius <= 113.645 or not, 7 samples belonging to the class SL are misclassified as class DH. Last but not least, the leaf node that the sacral slope <= 40.91 or not, 7 samples belonging to the class DH are misclassified as class SL.