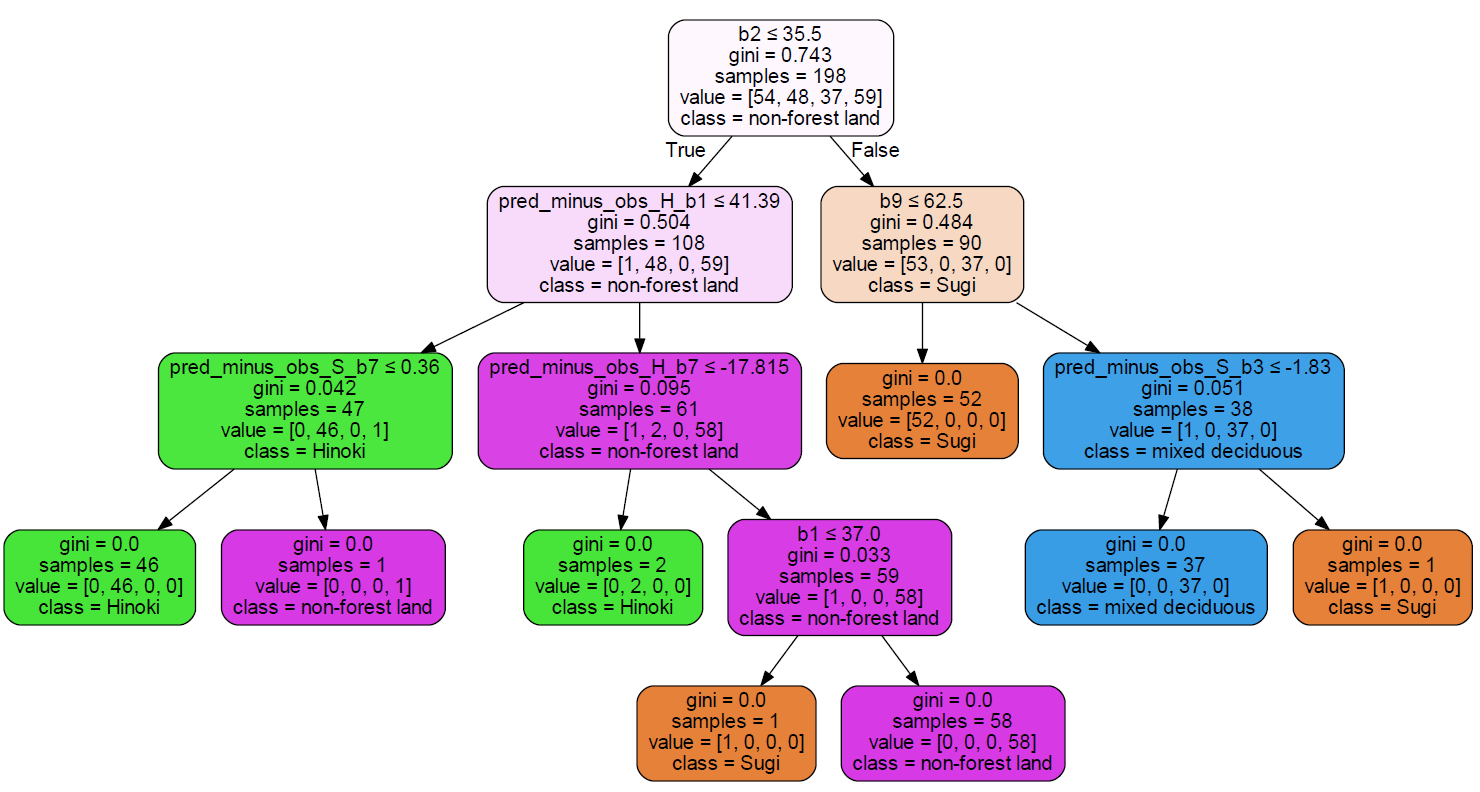
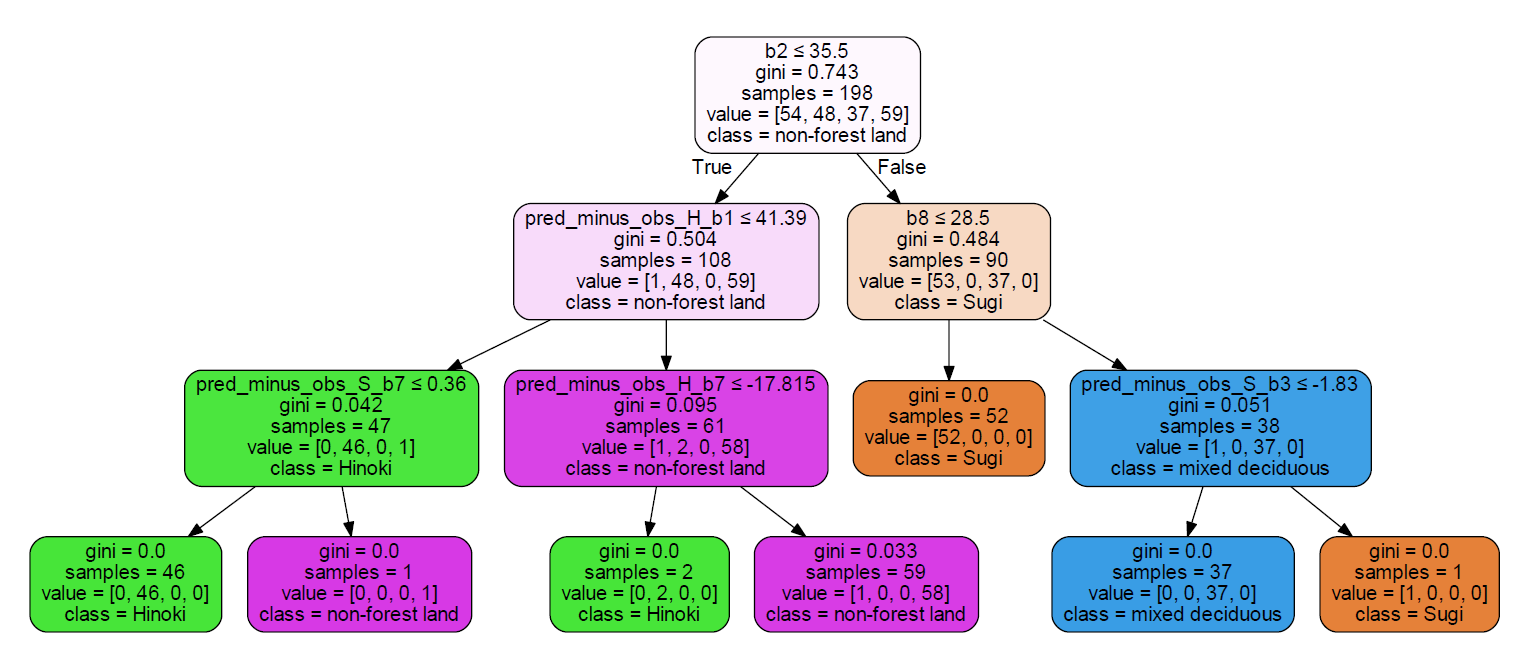
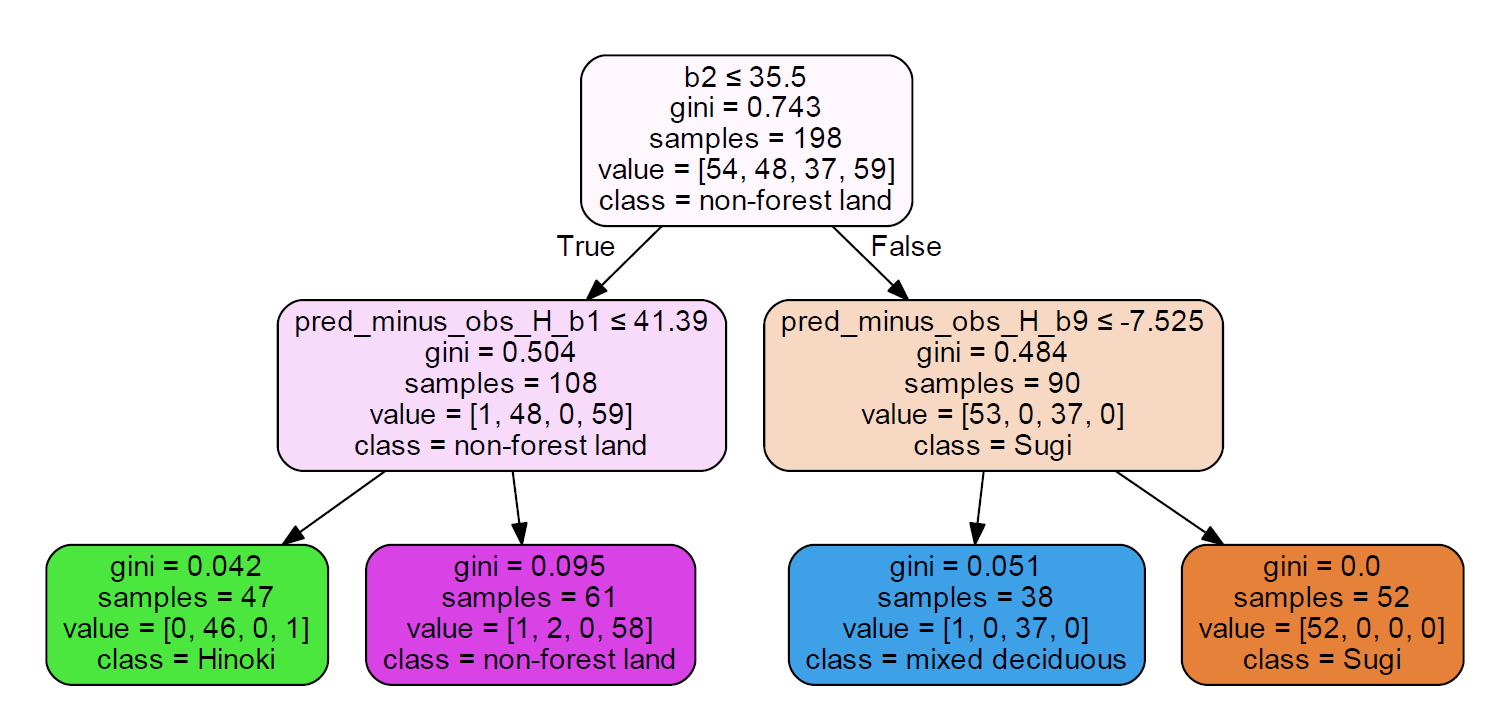
1. Training set partition

*Image 1: Training set using max depth =4*



*Image 2: Training set using max depth =3*



*Image 3: Training set using max depth =2*

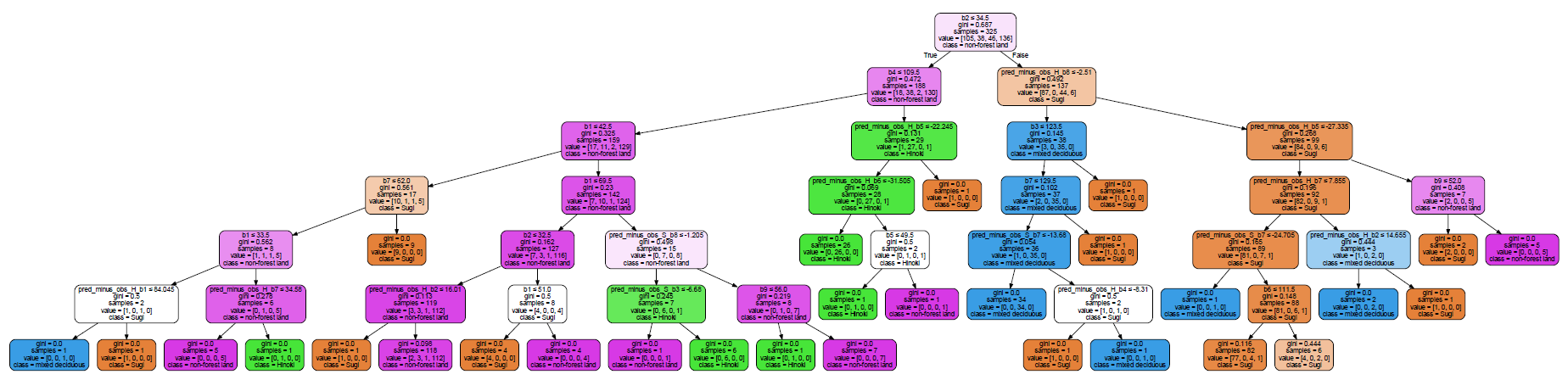
3 decision trees have been generated based on the training set data with 198 samples using the max depth as parameter. Firstly, we can see that the higher the depth, the deeper the tree, the more leaf nodes it has and it includes more information. Focusing on the zero Gini index, the higher depth has more zero Gini index, which summarized as below:

* 4 Max depths 🡪 8 zero Gini index
* 3 Max depths 🡪 6 zero Gini index
* 2 Max depths 🡪 1 zero Gini index

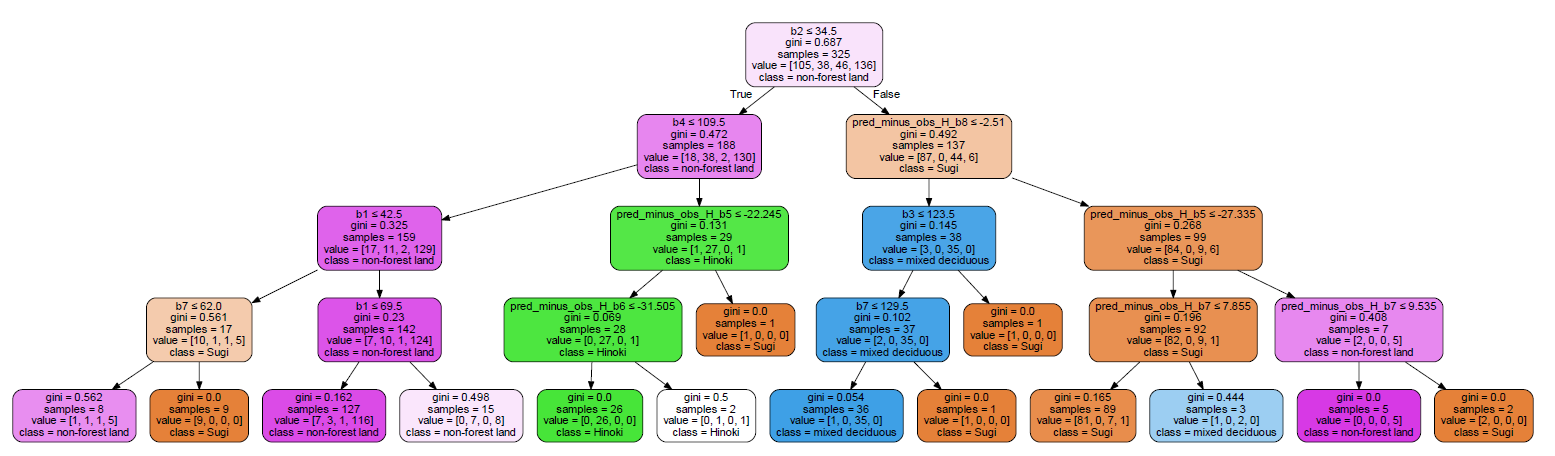
As a result, the higher the depth, the information should be more reliable as there are more zero Gini index. In the case of max depths equal of 4, refers to Image 1, Gini index of every end nodes of the tree is equal to zero, all 198 samples should be correctly classified. On the other hand, for the case of 2 max depths, refers to Image 3, there are only 1 end nodes has the zero Gini index, which contains of 52 samples, meaning that only 52 samples among 198 samples should belong to Sugi forest type, but for the other 3 classes, outliers occurred, for example 1 outlier which supposed to be non-forest land occured in Hinoki class among the 46 samples.

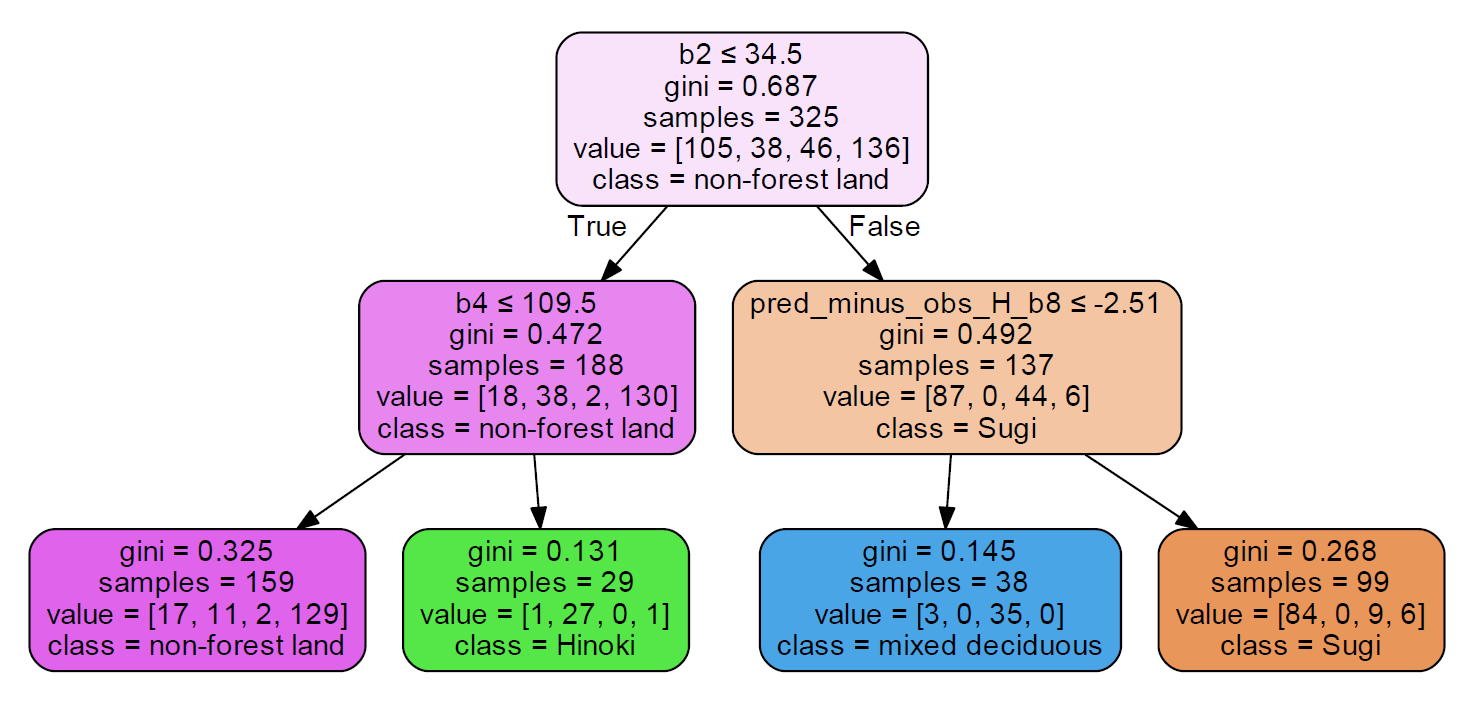
Training using training set contains sample of 198 and max depth equal to 4, and we got the accuracy rate of 0.797. Training using training set max depth equal to 2, and we got the accuracy rate of 0.8. Noted that the accuracy rate increased compare to the previous case when max depth equal to 4. This is because of overfitting, when the tree is larger, the test error begins to increase.

Since the data itself contains impurity, the decision tree with deeper maxdepth doesn’t seems to be aiding the classification. It is just a hypothesis for the classification, deeper decision tree doesn’t necessary support the classification better as the training data contain impurities. We should use the testing data to determine the performnace of our decision tree and determine the most appropriate depth.

(b) Testing set partition

*Image 4:* *Testing set using max depth =6*

*Image 5:* *Testing set using max depth =4*



*Image 6:* *Testing set using max depth =2*

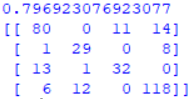
In the case of using testing set which contains 325 samples, we can see that even using the same parameter which is max depth equal to 4, the trees became more complicated, refers to Image 5. the This can be explained because more data sample, more noise will occur. Summary of the zero Gini index obtained from this testing data set is as below:

* 6 Max depths 🡪 26 zero Gini index
* 4 Max depths 🡪 7 zero Gini index
* 2 Max depths 🡪 0 zero Gini index

Although the depth had increased to 6 and there are 26 zero Gini index, but there are just only of 172 samples among 325 samples should be correctly classified. Also, many unnecessary nodes appeared when the depth number is higher.

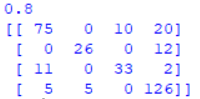
Training using testing set contains sample of 325 and max depth equal to 4, and we got the accuracy rate of 0.867. Training using testing set max depth equal to 2, and we got the accuracy rate of 0.894. Noted that the accuracy rate increased compare to the previous case when max depth equal to 4. As mentioned above, this is because of overfitting, when the tree is larger, the test error begins to increase.

(c) Classification performance and Confusion Matrix



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Predicted Class | | | |
| s | h | d | o |
| Actual Class | s | 80 | 0 | 11 | 14 |
| h | 1 | 29 | 0 | 8 |
| d | 13 | 1 | 32 | 0 |
| o | 6 | 12 | 0 | 118 |

*Table 1: Training set Max depth =4*

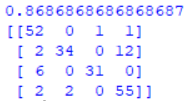
We can see that Sugi and mixed deciduous type are more likely to be confused with each other in this case, 13 samples were predicted to be Sugi but they should be mixed deciduous.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Predicted Class | | | |
| s | h | d | o |
| Actual Class | s | 75 | 0 | 10 | 20 |
| h | 0 | 26 | 0 | 12 |
| d | 11 | 0 | 33 | 2 |
| o | 5 | 5 | 0 | 126 |

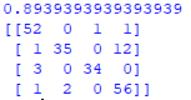
*Table 2: Training set Max depth =2*

Same with the upper case, Sugi and mixed deciduous type are more likely to be confused with each other, 11 samples were predicted to be Sugi but they should be mixed deciduous.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Predicted Class | | | |
| s | h | d | o |
| Actual Class | s | 52 | 0 | 1 | 1 |
| h | 2 | 34 | 0 | 12 |
| d | 6 | 0 | 31 | 0 |
| o | 2 | 2 | 0 | 55 |

*Table 3: Testing set Max depth =4*

We can see that Hinoki and non-forest type are more likely to be confused with each other in this case, 12 samples were predicted to be non-forest but they should be Hinoki forest.

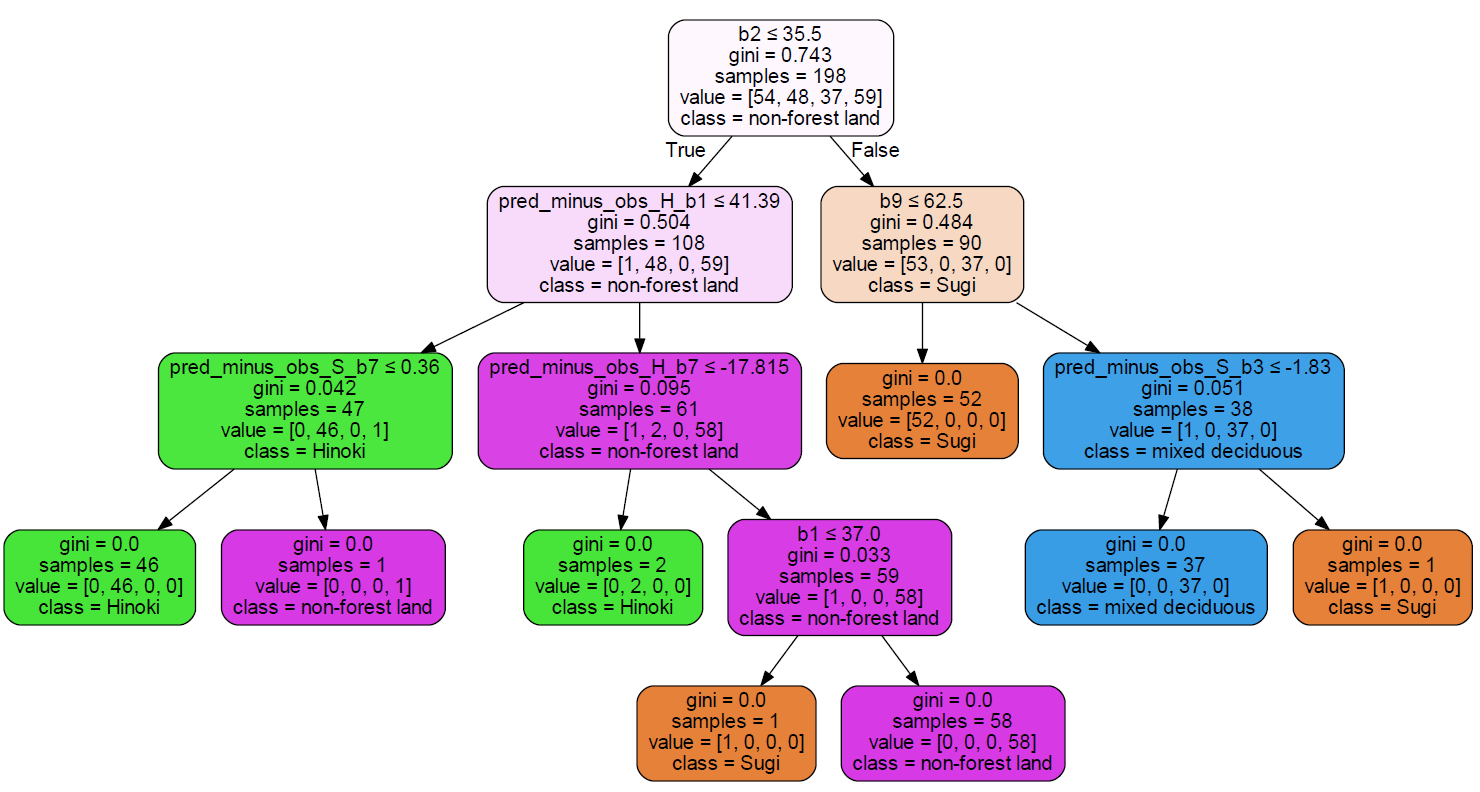


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Predicted Class | | | |
| s | h | d | o |
| Actual Class | s | 52 | 0 | 1 | 1 |
| h | 1 | 35 | 0 | 12 |
| d | 3 | 0 | 34 | 0 |
| o | 1 | 2 | 0 | 55 |

*Table 4: Testing set Max depth =2*

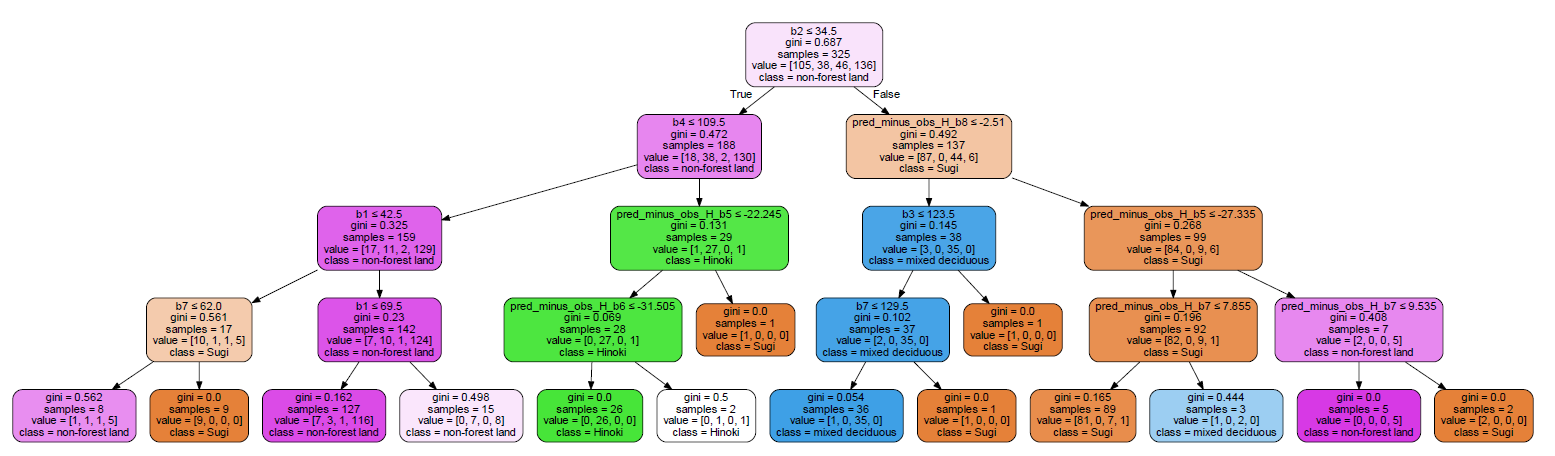
Same with the upper case, Hinoki and non-forest type are more likely to be confused with each other, 12 samples were predicted to be non-forest but they should be Hinoki forest.

To conclude, compare with the training set and testing set, overall accuracy rate of testing set is higher than training test, this can be explained because of training set has lesser sample data, and by using lesser sample data to train, the result might be too general so the result is not that reliable than testing set.

(d)

*Image 1: Training set using max depth =4*

As mentioned above, Sugi and mixed deciduous type are more likely to be confused with each other. The leaf node of pred\_minus\_obs\_S\_b3 ≤ -1.83 split the class of Sugi and mixed deciduous, so this node might lead to misclassification. Also, noted that variable b9 ≤ 62.5 also the classification between Sugi and mixed deciduous, in this case, since there are only 198 samples, all of the samples were correctly classify, but if the sample data is large, misclassification may occur.

*Image 5:* *Testing set using max depth =4*

As mentioned above, Hinoki and non-forest type are more likely to be confused with each other. The leaf node of b4 ≤ 109.5 split the class of Hinoki and non-forest, so this node might lead to misclassification.