City University of Hong Kong
2024 - 2025 Semester B
CS3481 Fundamentals of Data Science

Decision Tree Assignment

Report

1. Introduction

In this assignment, we construct multiple decision trees using the Wine dataset to classify wine samples into three distinct classes. Each decision tree is configured with different train-test split ratios, and keeps the impurity measures constant. The models are evaluated by comparing their tree structure, classification accuracy, and confusion matrices to identify patterns in misclassification.

2. Dataset Description

- **Features**
  - 13 chemical attributes (e.g., alcohol, malic acid, flavonoids).
- **Classes**
  - 3 types of wines (Class 0, 1, 2).
- **Class Distribution**
  - Class 0: 59 instances
  - Class 1: 71 instances
  - Class 2: 48 instances
  - Total: 178 instances

3. Results and Analysis

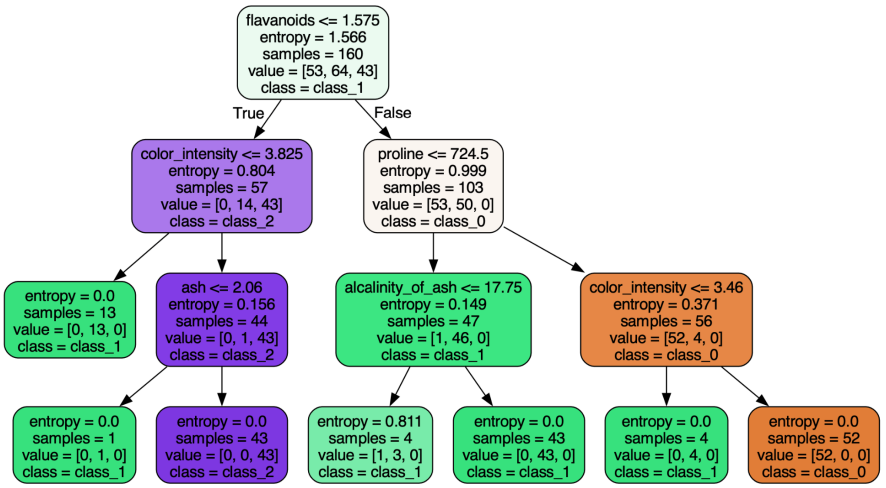
3.1. Task (a) & (b): Construction and Comparison of Multiple Different Decision Trees

In this task, the dataset was partitioned into training and test sets with ratios of 9:1, 8:2, 7:3, and 6:4. The 4 decision trees were built using the entropy impurity measure, which evaluates the information gain at each split. The maximum depth of the trees is 3, preventing overfitting and to maintain interpretability.

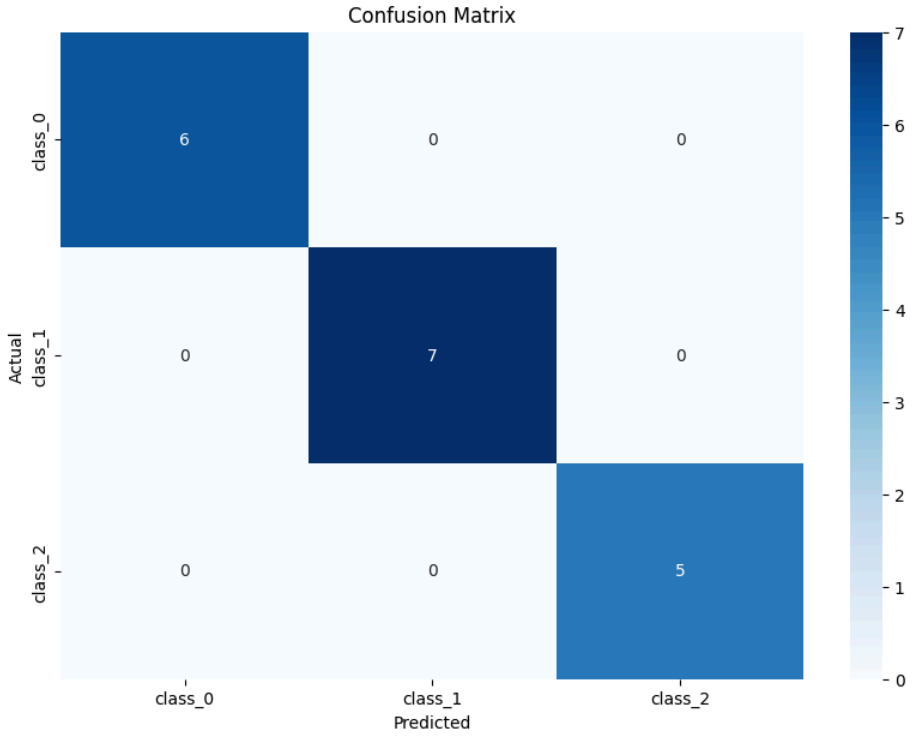
Tree configuration:

Tree	Partition Ratio (Train:Test)	Criterion	Max Depth
1	9:1	entropy	3
2	8:2	entropy	3
3	7:3	entropy	3
4	6:4	entropy	3

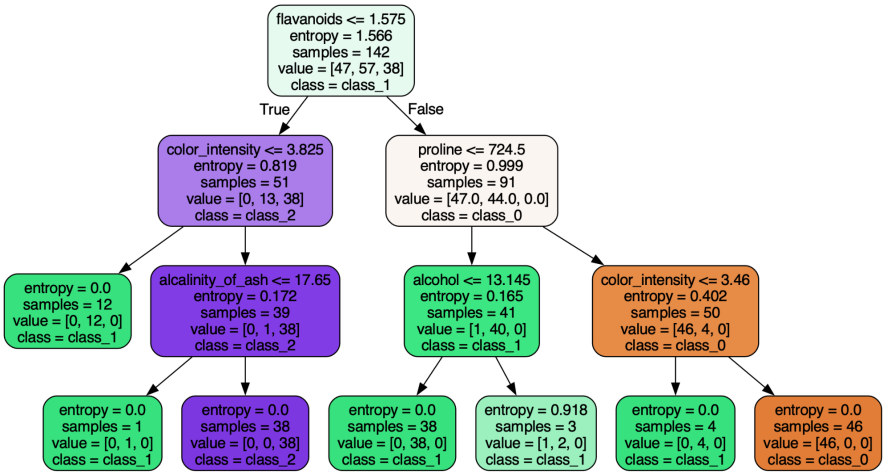
Tree 1 with partition ratio 9:1:



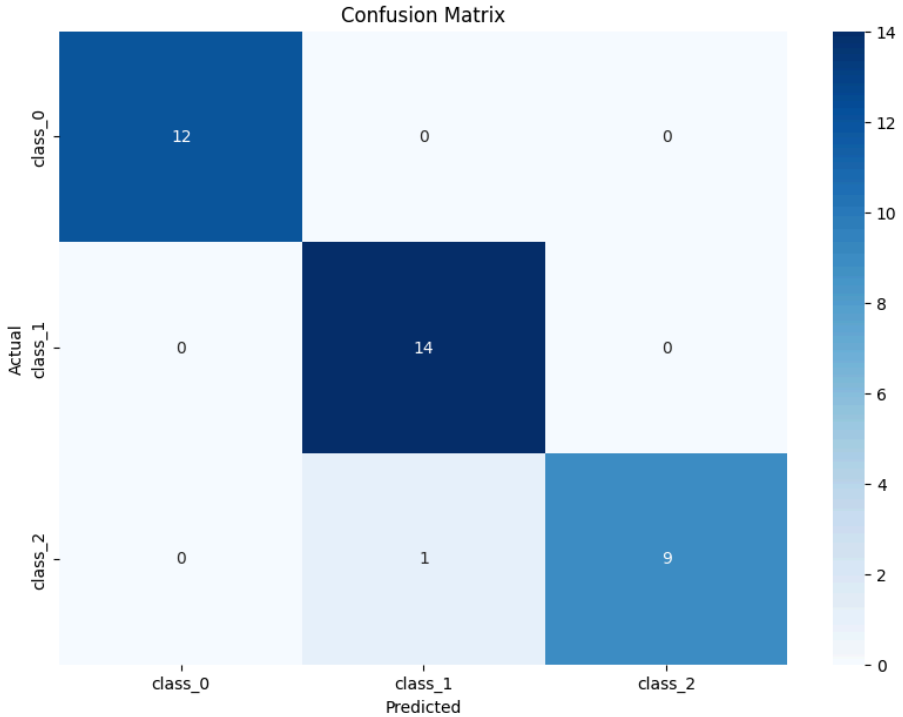

```
Classification Report:
              precision    recall  f1-score   support

     class_0       1.00      1.00      1.00         6
     class_1       1.00      1.00      1.00         7
     class_2       1.00      1.00      1.00         5

    accuracy                           1.00        18
   macro avg       1.00      1.00      1.00        18
weighted avg       1.00      1.00      1.00        18

Training set accuracy: 0.993750
Testing set accuracy: 1.000000
```



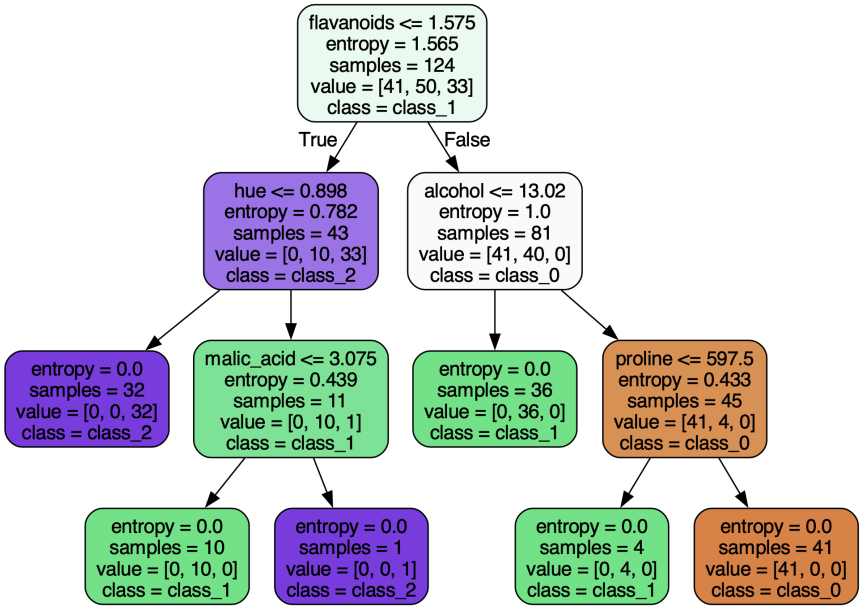
Tree 2 with partition ratio 8:2:



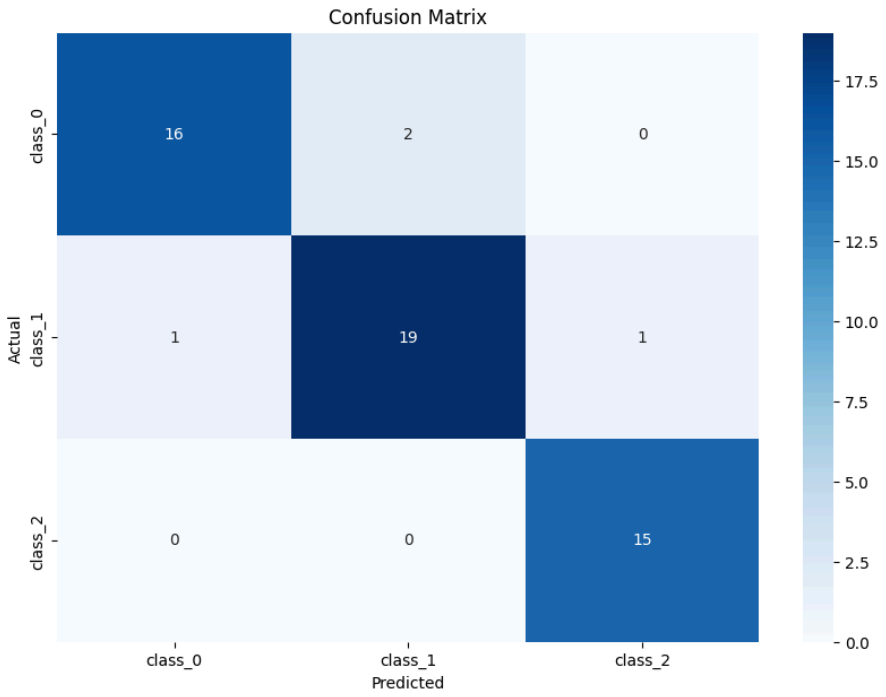

```
Classification Report:
              precision    recall  f1-score   support

     class_0       1.00      1.00      1.00        12
     class_1       0.93      1.00      0.97        14
     class_2       1.00      0.90      0.95        10

    accuracy                           0.97        36
   macro avg       0.98      0.97      0.97        36
weighted avg       0.97      0.97      0.97        36

Training set accuracy: 0.992958
Testing set accuracy: 0.972222
```



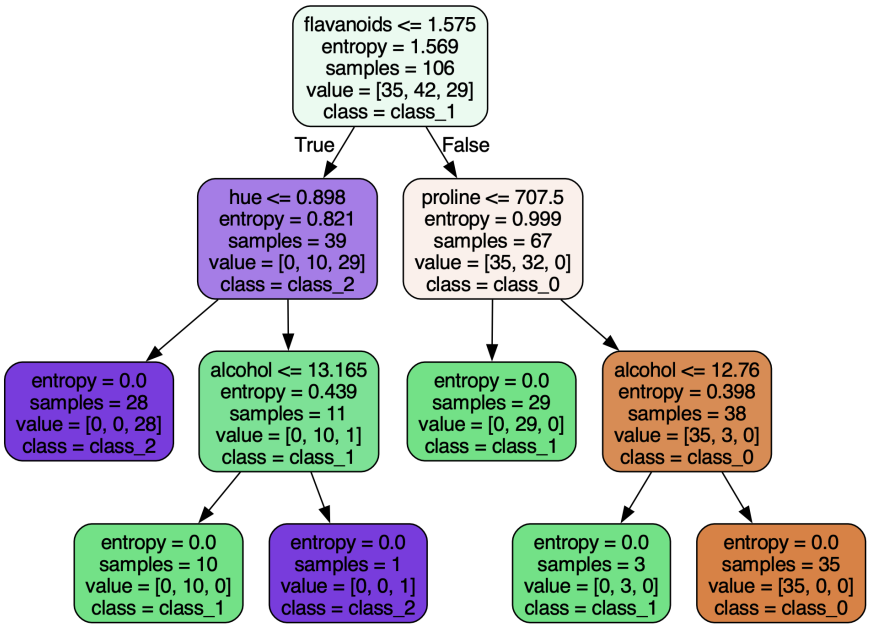
Tree 3 with partition ratio 7:3:



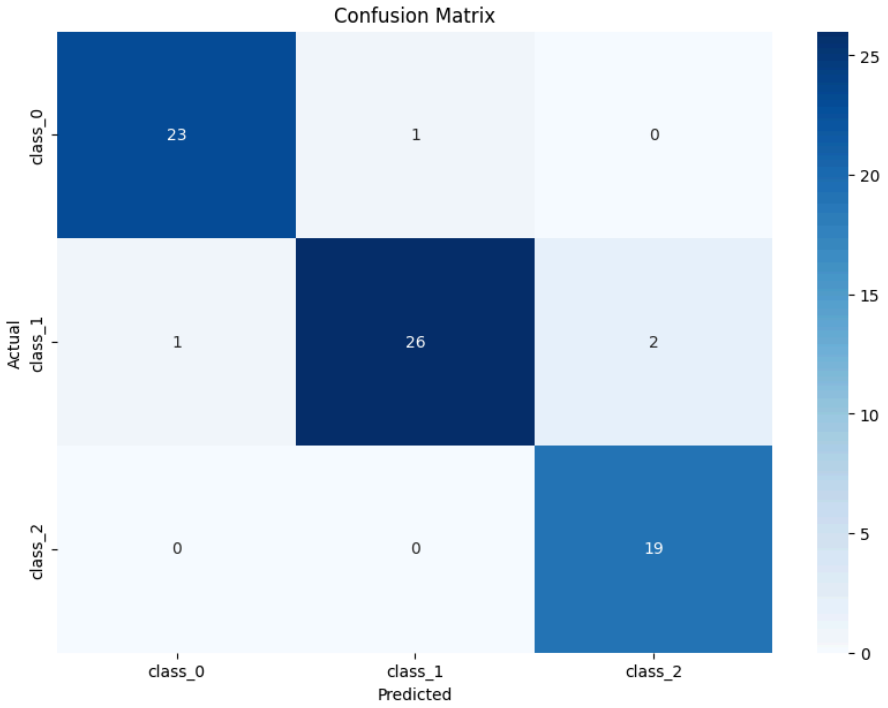

```
Classification Report:
              precision    recall  f1-score   support

     class_0       0.94      0.89      0.91        18
     class_1       0.90      0.90      0.90        21
     class_2       0.94      1.00      0.97        15

    accuracy                           0.93        54
   macro avg       0.93      0.93      0.93        54
weighted avg       0.93      0.93      0.93        54

Training set accuracy: 1.000000
Testing set accuracy: 0.925926
```



Tree 4 with partition ratio 6:4:




```
Classification Report:
              precision    recall  f1-score   support

     class_0       0.96      0.96      0.96        24
     class_1       0.96      0.90      0.93        29
     class_2       0.90      1.00      0.95        19

    accuracy                           0.94        72
   macro avg       0.94      0.95      0.95        72
weighted avg       0.95      0.94      0.94        72

Training set accuracy: 1.000000
Testing set accuracy: 0.944444
```



Comparing the tree structure of the 4 trees:

Tree 1 and Tree 2 have a total of 7 leaf nodes, which is more than the other 2 trees with 5 leaf nodes. The more leaf nodes in Tree 1 and Tree 2 may indicate a potential overfitting risk, while being too specific for the training data. One of the evidence is that there are some leaf nodes contain only one training sample.

All the trees are starting by splitting the samples with the boundary “flavonoids = 1.575”, but become different on the children of the root node. Also, all Class_0 and Class_2 are splitting well by the root node.

The accuracy of the trees:

Tree	Partition Ratio (Train:Test)	Accuracy(Train)	Accuracy(Test)
1	9:1	0.993750	1.000000
2	8:2	0.992958	0.972222
3	7:3	1.000000	0.925926
4	6:4	1.000000	0.944444

Tree 1 has 100% accuracy on the test set, which is the highest test set accuracy among 4 trees. But its reliability is limited due to the extremely small test set (18 samples). Compared to this unrealistic prediction model with 100% accuracy, larger partitions (e.g. 8:2 or 6:4) show slight performance declines (94-97% accuracy), reflecting more realistic generalization. In this case, the perfect scores on testing may stem from a non-representative test subset rather than true robustness. In other words, it could just be luck to achieve a perfect score.

In contrast, Trees 3 and 4 exhibit significant overfitting. Although their training accuracy is perfect (100%), their test performance is lower (92.6% and 94.4%, respectively). This large gap between training and test accuracy indicates that these models memorize noise in the training data, leading to erratic generalization on unseen examples. For instance, Tree 3 shows degraded recall for Class_0 (89%), while Tree 4 has misclassifications on Class_1, reflecting overlapping feature distributions in the dataset.

Tree 2 performs an optimal balance by using 80% of the data for training and 20% for testing. It has extremely high training accuracy (99.3%) and high test accuracy (97.2%), where the gap between training and test accuracy is only around 2%. It suggests a model that learns meaningful patterns without overfitting. The moderately sized test set (36 samples) reduces variance and provides a trustworthy evaluation, unlike the statistically unreliable 9:1 split. Additionally, Tree 2 maintains consistent and high precision and recall across all classes (≥93%), avoiding the class-specific instability seen in Trees 3 and Tree 4.

3.2. Task (c): Finding Out Confused Class Pairs

This task is to identify confused class pairs from the 4 trees selected before. Since Tree 1 has perfect accuracy on testing and no misclassification exists, we skip the analysis of Tree 1 in this task.

For Tree 2 with partition ratio 8:2:

		Predicted Class		
		Class_0	Class_1	Class_2
Actual Class	Class_0	12	0	0
	Class_1	0	14	0
	Class_2	0	1	9

There is only 1 misclassification in the tree, which a Class_2 sample misclassified as Class_1. It seems a specific subset of Class_2 overlaps with Class_1 in feature space, leading to isolated misclassification. Therefore, Class_1 and Class_2 are most likely to be confused with each other from this tree.

Confusion pair:

- Class_1 and Class_2

For Tree 3 with partition ratio 7:3:

		Predicted Class		
		Class_0	Class_1	Class_2
Actual Class	Class_0	16	2	0
	Class_1	1	19	1
	Class_2	0	0	15

There are few misclassification in the tree:

- 2 Class_0 sample misclassified as Class_1
- 1 Class_1 sample misclassified as Class_0
- 1 Class_1 sample misclassified as Class_2

Classification error rate for each class:

Class	Error rate
Class_0	2/18 = 0.1111
Class_1	2/21 = 0.0952
Class_2	0/15 = 0.0000

We can see Class_0 has the highest classification error rate, and all errors come from misclassified as Class_1. Also, there is one Class_1 sample misclassified as Class_0. Therefore, Class_1 and Class_2 are most likely to be confused with each other from this tree.

There is also 1 misclassification in the tree, which a Class_1 sample misclassified as Class_2. It seems a specific subset of Class_1 overlaps with Class_2 in feature space, leading to isolated misclassification. Therefore, Class_1 and Class_2 are also a confusion pair from the tree.

Confusion pair:

- Class_0 and Class_1 (most confused pair)
- Class_1 and Class_2

For Tree 4 with partition ratio 6:4:

		Predicted Class		
		Class_0	Class_1	Class_2
Actual Class	Class_0	23	1	0
	Class_1	1	26	2
	Class_2	0	0	19

There are few misclassification in the tree:

- 1 Class_0 sample misclassified as Class_1
- 1 Class_1 sample misclassified as Class_0
- 2 Class_1 sample misclassified as Class_2

Classification error rate for each class:

Class	Error rate
Class_0	1/24 = 0.0417
Class_1	3/29 = 0.1034
Class_2	0/19 = 0.0000

We can see Class_1 has the highest classification error rate, and 2 errors come from misclassified as Class_2. It seems a specific subset of Class_1 overlaps with Class_2 in feature space, leading to isolated misclassification. Therefore, Class_1 and Class_2 most likely to be a confusion pair from the tree.

Since both Class_0 and Class_1 have a sample misclassified as each other, Class_0 and Class_1 are also a confusion pair from the tree.

Confusion pair:

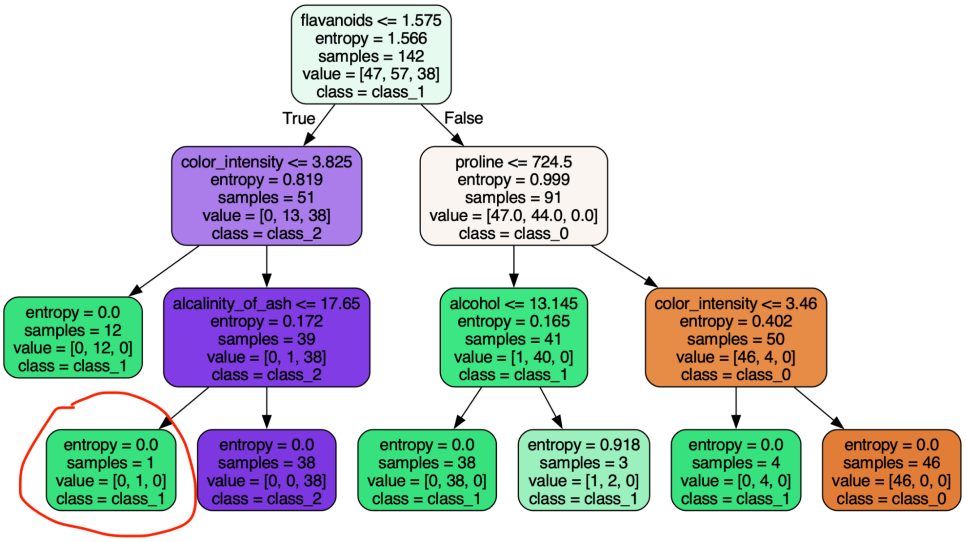
- Class_1 and Class_2 (most confused pair)
- Class_0 and Class_1

3.3. Task (d): Analyzing Leaf Nodes and Decision Paths

For the confused class pairs identified in Task (c), this task is to locate the leaf nodes responsible for misclassifications, trace the decision paths leading to these errors, and analyze why the model struggles to distinguish between the classes.

For Tree 2 with partition ratio 8:2:

The decision sequence of the only misclassified sample in depth-first order:
[1 1 0 1 1 0 0 0 0 0 0 0 0]

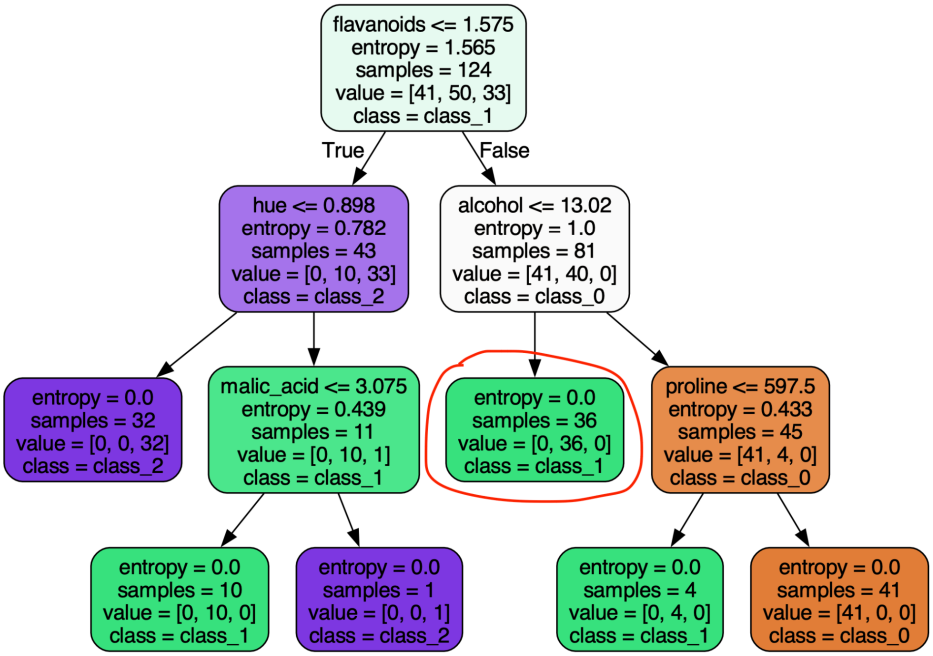


We can see the misclassified sample has feature values with flavanoids <= 1.575, color_intensity <= 3.825 and alcalinity_of_ash <= 17.65. This misclassified Class_2 sample likely shares similar alcalinity_of_ash values with the one Class_1 sample in the training set, leading to the misclassification.

For Tree 3 with partition ratio 7:3:

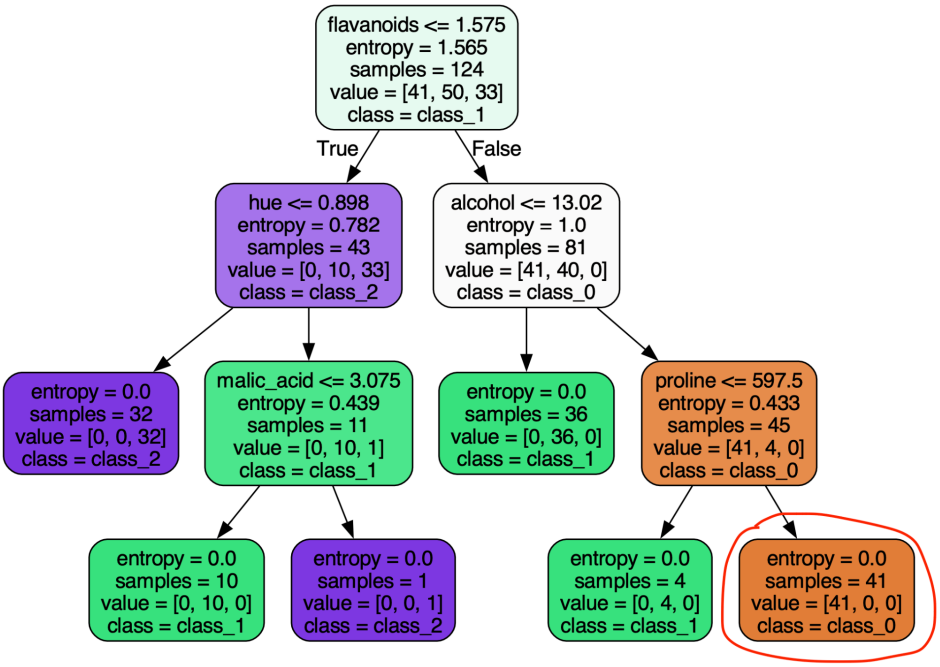
The decision sequence (in depth-first order) of the two Class_0 sample, which misclassified as Class_1:

[1 0 0 0 0 0 1 1 0 0 0]



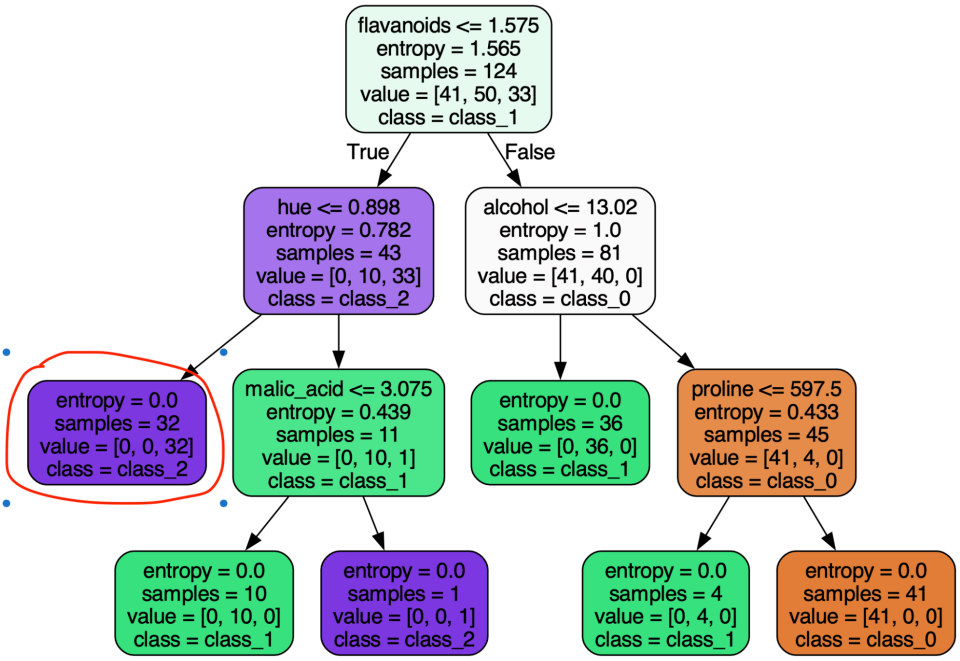
We can see the two misclassified samples have feature values with flavanoids > 1.575 and alcohol <= 13.02, which shares similar alcohol values with the most Class_1 sample in the training set, leading to the misclassification.

The decision sequence (in depth-first order) of a Class_1 sample, which misclassified as Class_0:
[1 0 0 0 0 0 1 0 1 0 1]



We can see the misclassified sample has feature values with flavanoids > 1.575, alcohol > 13.02 and proline > 597.5, which shares similar proline values with all Class_0 samples in the training set, leading to the misclassification.

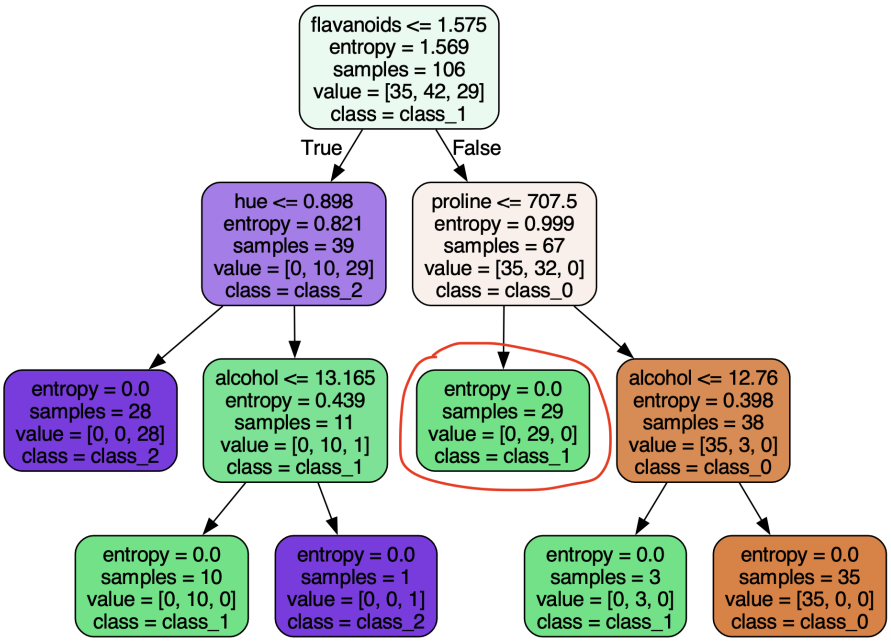
The decision sequence (in depth-first order) of a Class_1 sample, which misclassified as Class_2:
[1 1 1 0 0 0 0 0 0 0 0]



We can see the misclassified sample has feature values with flavanoids <= 1.575 and hue <= 0.898, which shares similar hue values with the most Class_2 samples in the training set, leading to the misclassification.

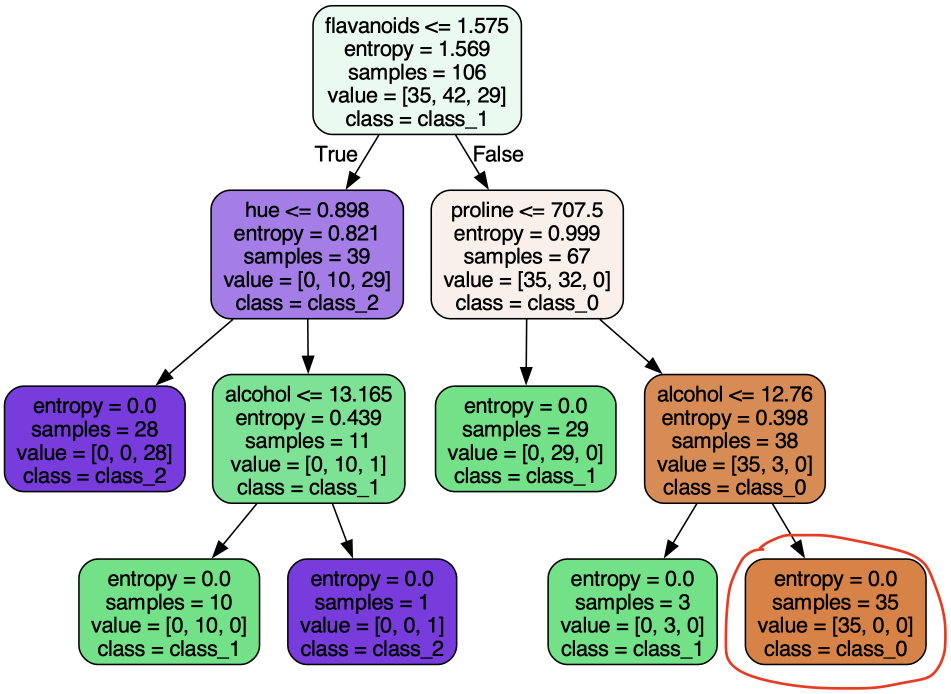
In Tree 3, it seems Class_0 and Class_1 share overlapping alcohol and proline values, and Class_1 and Class_2 share overlapping hue values.

For Tree 4 with partition ratio 6:4:
The decision sequence (in depth-first order) of a Class_0 sample, which misclassified as Class_1:
[1 0 0 0 0 0 1 1 0 0 0]



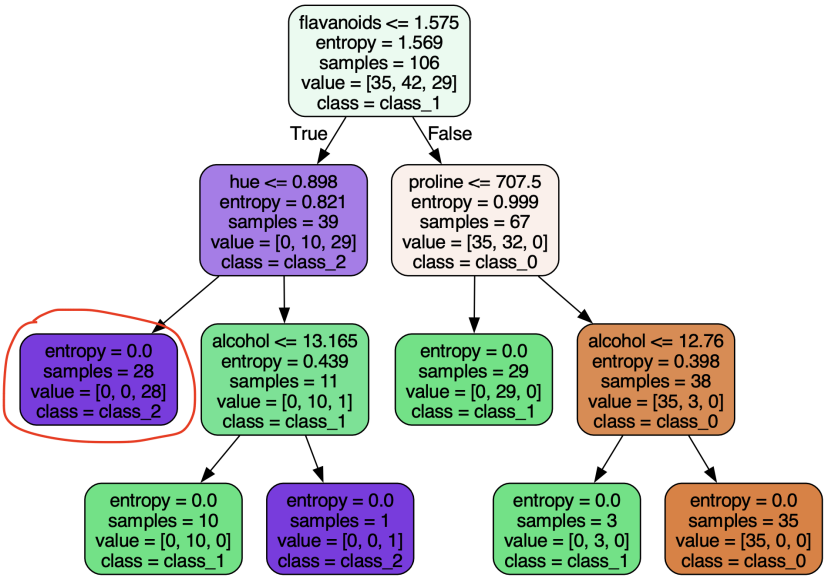
We can see the misclassified sample has feature values with flavanoids > 1.575 and proline <= 707.5, which shares similar proline values with a subset of Class_1 samples in the training set, leading to the misclassification.

The decision sequence (in depth-first order) of a Class_1 sample, which misclassified as Class_0:
[1 0 0 0 0 0 1 0 1 0 1]



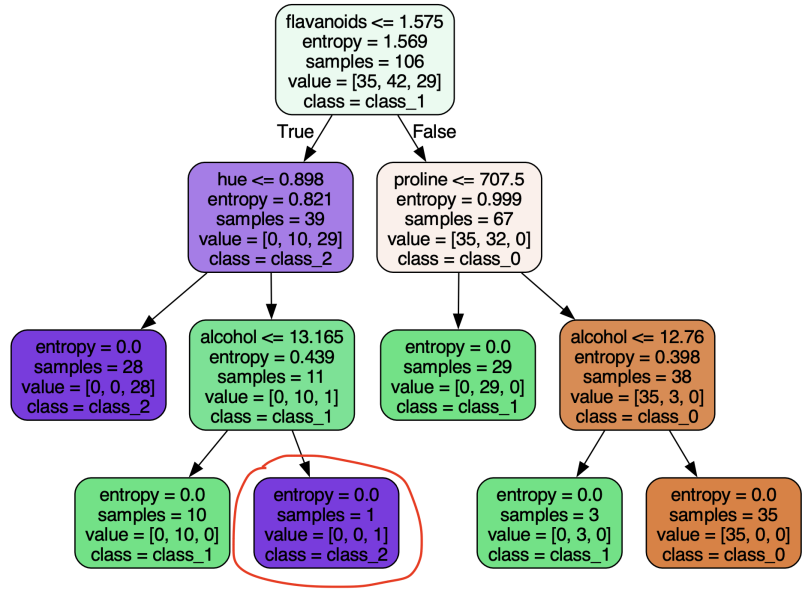
We can see the misclassified sample has feature values with flavanoids > 1.575, proline > 707.5 and alcohol > 12.76, which shares similar alcohol values with all Class_0 samples in the training set, leading to the misclassification.

There are two Class_1 samples which are misclassified as Class_2 in Tree 4.
The decision sequence (in depth-first order) of one of them:
[1 1 1 0 0 0 0 0 0 0 0]



We can see the misclassified sample has feature values with flavanoids <= 1.575 and hue <= 0.898, which shares similar hue values with the most Class_2 samples in the training set, leading to the misclassification.

The decision sequence (in depth-first order) of another one:
[1 1 1 0 0 0 0 0 0 0 0]



We can see the misclassified sample has feature values with flavanoids <= 1.575, hue <= 0.898 and alcohol > 13.165, which shares similar alcohol values with a Class_2 sample in the training set, leading to the misclassification.

In Tree 4, it seems Class_0 and Class_1 share overlapping alcohol and proline values, and Class_1 and Class_2 share overlapping hue and alcohol values.