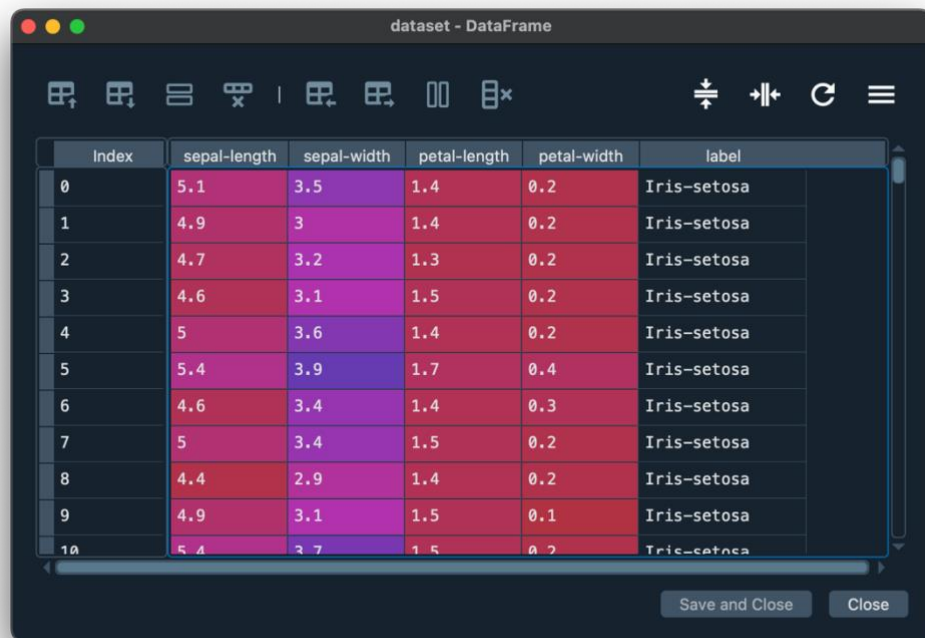


### Part 2 Questions

**2.1.** The dimensions of the dataset are attributes related to the Iris plant such as sepal-length, sepal-width, petal-length, petal-width, and the target labels for classification. Excluding the index column, there are 5 features (columns) with 150 instances (data points).



Index	sepal-length	sepal-width	petal-length	petal-width	label
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa

**2.2.** The labels of the dataset is what the decision tree classifier is trying predict or the targets, the type of Iris plant.

**2.3.** Considering the 3 ways the dataset was partitioned; the predictions came out with a very consistently discrete set of accuracy scores after repeated iterations of training and predictions. The size of training sets, of which the complement was the testing set, was linearly proportional to the accuracy score of the predicted results. The following table summarizes the results:

Partition Option	Training Set Size	Test Set Size	Average Prediction Accuracy Score (10 Iterations)	Average Correct Predictions	Average Wrong Predictions
A	75	75	33.3%	25	50
B	100	50	0.0%	0	50
C	125	25	84.0%	22	3

With the C partition option, the size of the training set being 125 instances, the prediction accuracy score had some variance, compared to the B or A partitioning options. On most individual prediction runs; the accuracy score came out either 80.0% or 88.0%. Ultimately, this could have been because of the way in which the dataset was partitioned with almost 83% being the training set, and only about 17% being the test set. Out of 5 times, the prediction percentage for the C partition option, 3 times were 80.0% and the rest 2 times were 88.0%.

**2.4.** Some splits affect the prediction accuracy score in a stronger manner because of the amount of data the model gets to train with. More data indicates a larger, but also a more balanced decision tree. This is because the model has been exposed to a truer representation of the dataset and so, it can make more accurate predictions on unfamiliar or unseen data. Minor changes in the size of the training set can drastically affect the accuracy of the predictions. Since this is not a binary classification but a ternary classification of Iris plants, the increase in complexity calls for a more diverse dataset for training. In this assignment, because the classes of the dataset were sequentially listed, lower partitions which indicate smaller training datasets, would have only 1 or 2 target labels, but a larger dataset can have all the target labels of the dataset, so it would result in a higher accuracy score with repeated training and predictions.

#### Output Images

```
Predicting Labels...

Results:
Testset Size      : 75
Correct Preditions: 25
Accuracy Percent  : 33.333%
```

```
Predicting Labels...

Results:
Testset Size      : 50
Correct Preditions: 0
Accuracy Percent  : 0.000%
```

```
Predicting Labels...

Results:
Testset Size      : 25
Correct Preditions: 20
Accuracy Percent  : 80.000%
```

```
Predicting Labels...

Results:
Testset Size      : 25
Correct Preditions: 22
Accuracy Percent  : 88.000%
```