Project 1. Machine Learning (STA 5365). Aditya Ranjan Bhattacharya (arb17b)

For Project 1 we used DecisionTreeClassifier model from sklearn[1] as our decision tree model. In this case, we used 2 separate functions data_to_numpy() and label_to_numpy(), which loads the features and labels from the files to a numpy array.

In order to get all the datasets running at the same time, we put the filenames of the training features, training labels, testing features and testing labels in 4 separate lists. The list with the features for both training and testing also contains the number of features for the dataset as the 2nd entry.

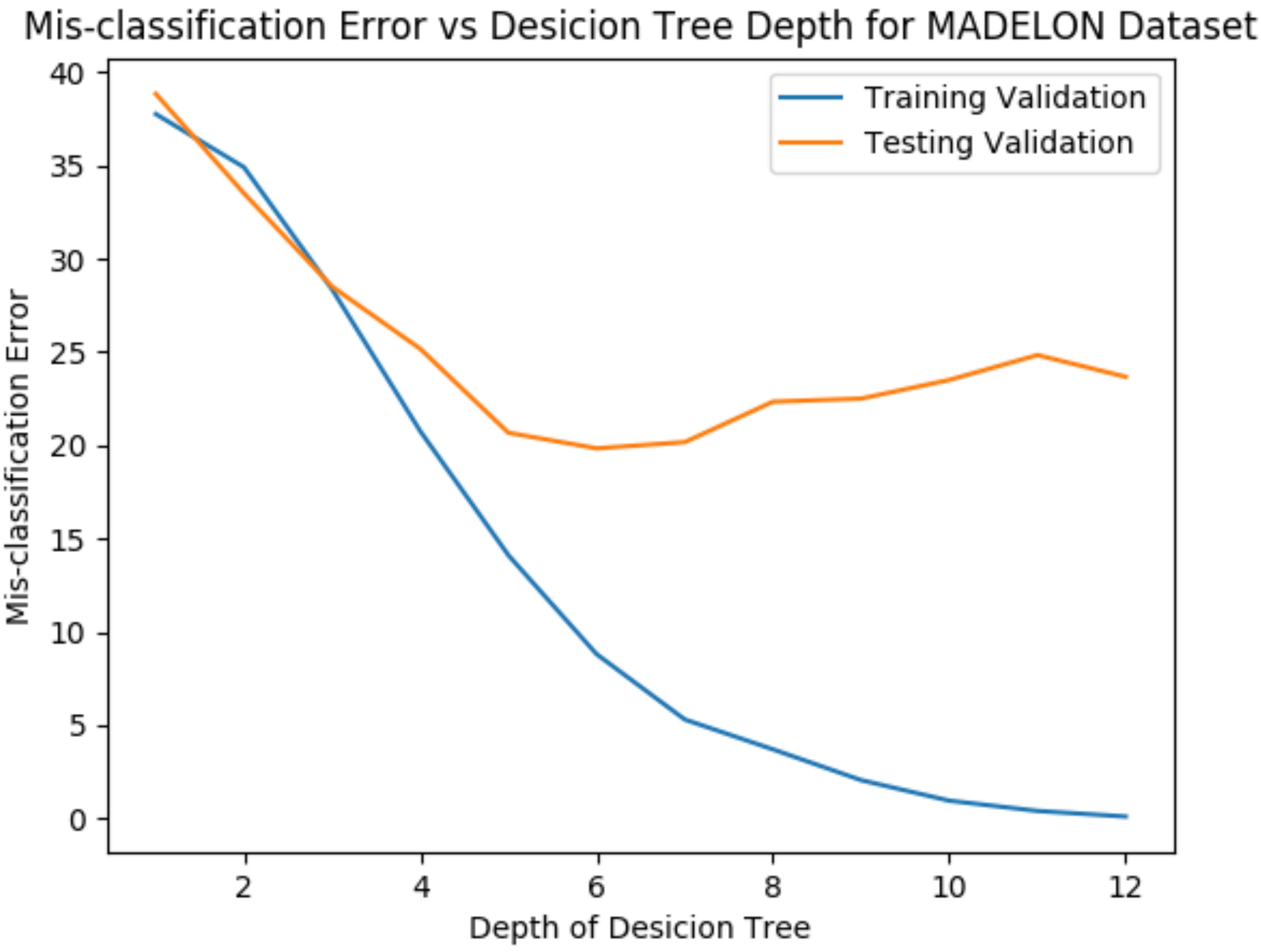
Once the datasets are loaded, the model is trained on the training set of the data, and then validation checks using both training and testing set are done. This happens iteratively, as the model goes from having depth = 1 to depth = 12. These are then saved in 2 separate numpy arrays, one for the training validation, and the other for testing validation.

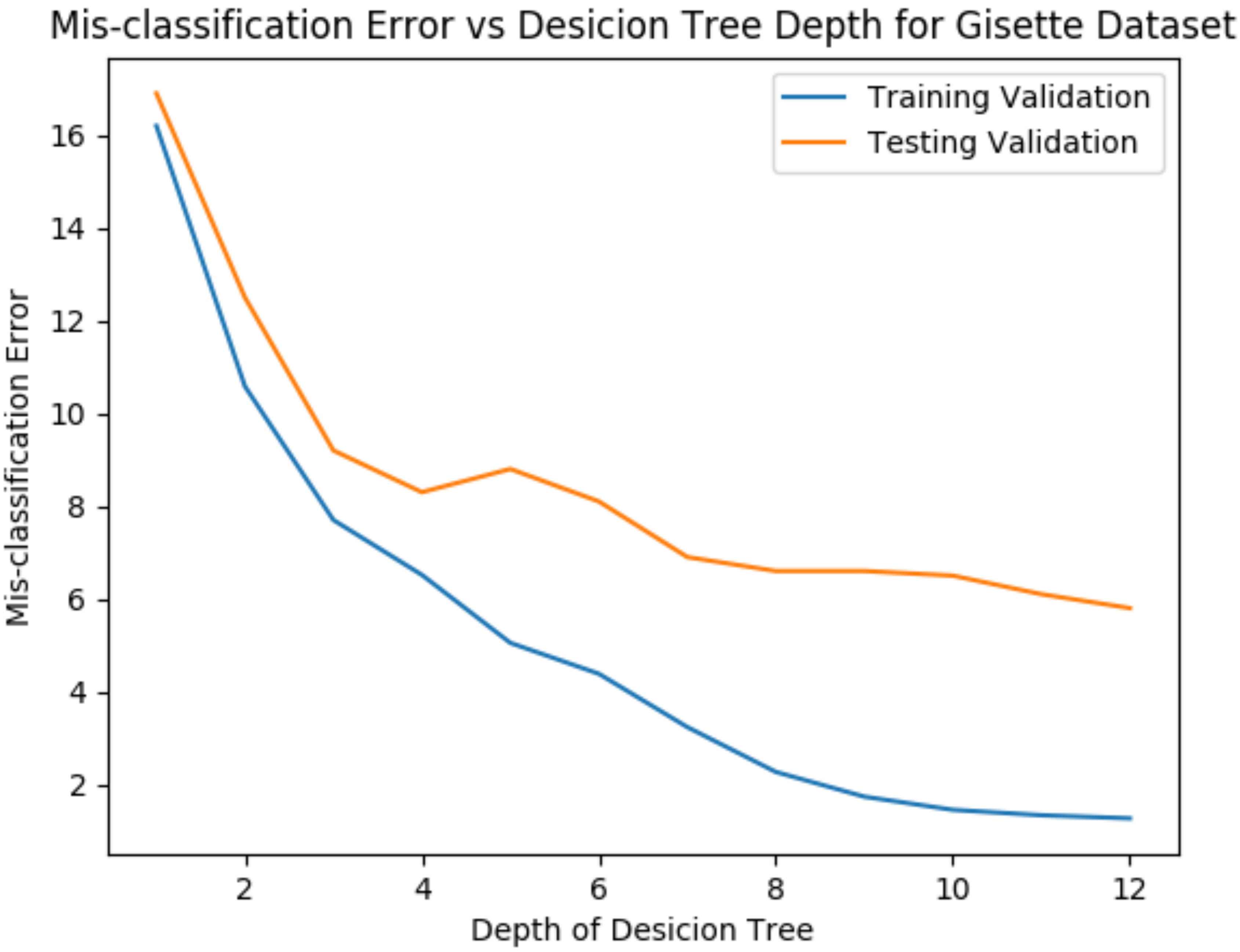
Once all 12 models' accuracy parameters are saved in the arrays, we plot them first, and then find the index of the testing validation with the lowest value of mis-classification error. Now as array indexing starts from 0, but the minimum depth of the decision tree model is 1. Thus decision tree model with depth = 1 has its mis-classification error saved in scores_test[0], and in general depth = x has its mis-classification error on testing set saved in scores_test[x-1].

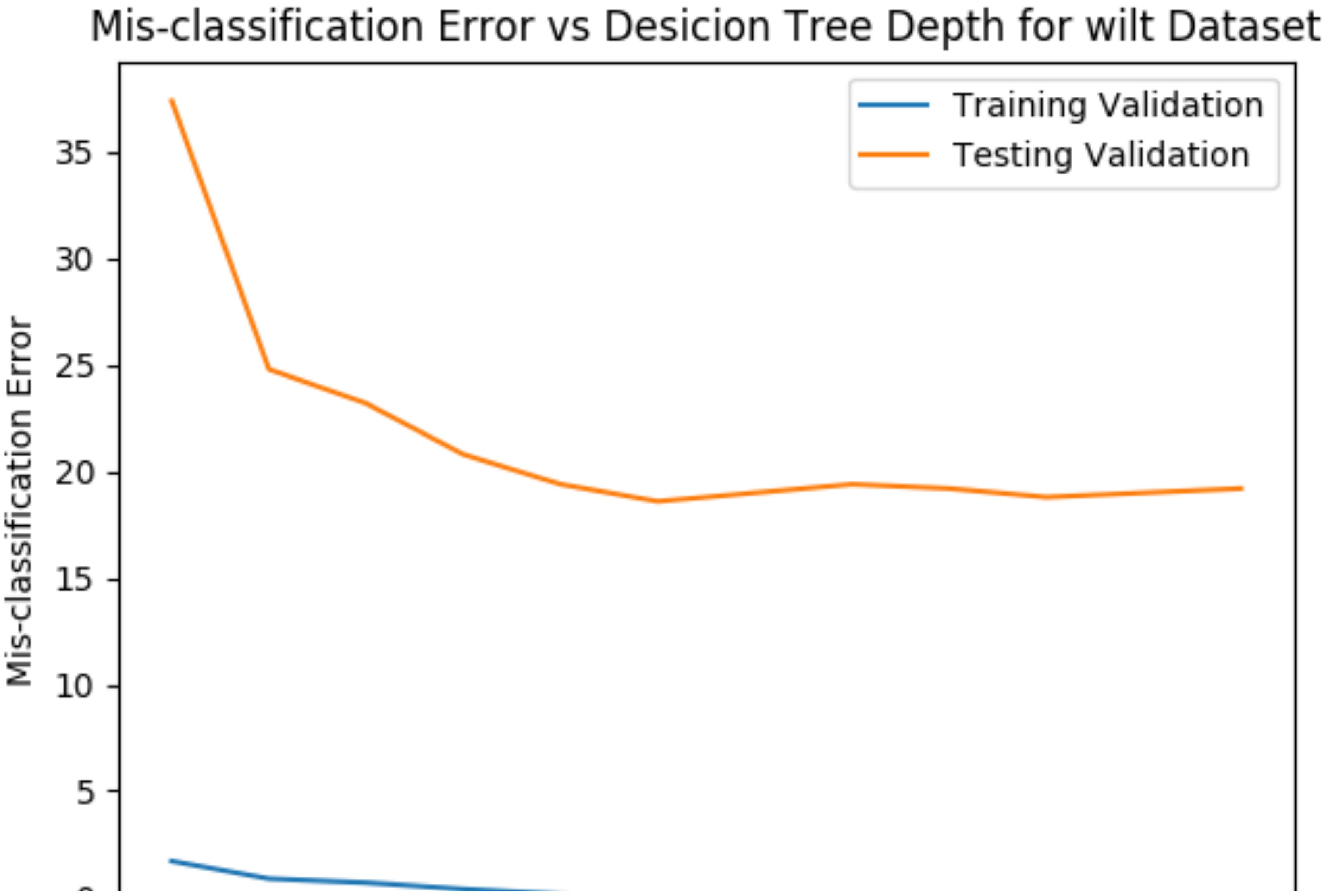
The results are given below.

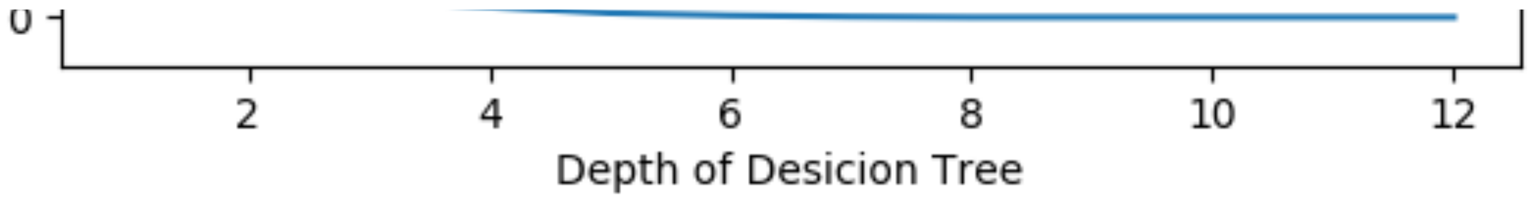
Min. Test Classification Error and Related Tree Depth For Datasets

Dataset	Depth of Decision Tree with Minimum Test Classification Error	Value of Minimum Test Classification Error (%)
Madelon	6	19.83
Gisette	12	5.80
Wilt	6	18.60




Mis-classification Error vs Desicion Tree Depth for Gisette Dataset
Training Validation
Testing Validation
Mis-classification Error
16
14
12
10
8
6
4
2
2
4
6
8
10
12
Depth of Desicion Tree


Mis-classification Error vs Desicion Tree Depth for wilt Dataset
Training Validation
Testing Validation
Mis-classification Error
35
30
25
20
15
10
5



[1] Lars Buitinck, Gilles Louppe, Mathieu Blondel, et. al, *“API design for machine learning software: experiences from the scikit-learn project”*, **ECML PKDD Workshop: Languages for Data Mining and Machine Learning** (2013, pages 108-122)