UNIVERSITY OF SCIENCE – VIET NAM NATIONAL UNIVERSITY OF HO CHI MINH CITY

REPORT

LAB 02: DECISION TREE WITH SCIKIT-LEARN

Full name: Nguyễn Hứa Hùng

ID: 19127150

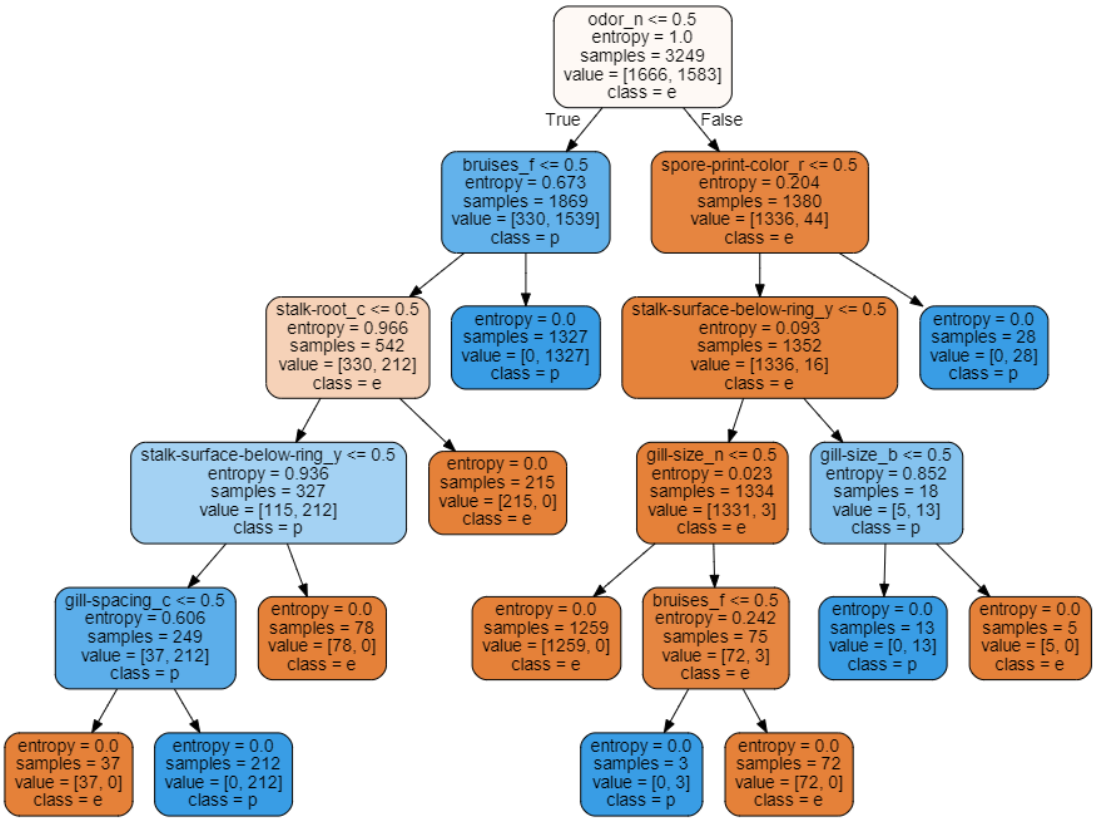
Class: 19CLC2

I. PREPARING THE DATA SETS:

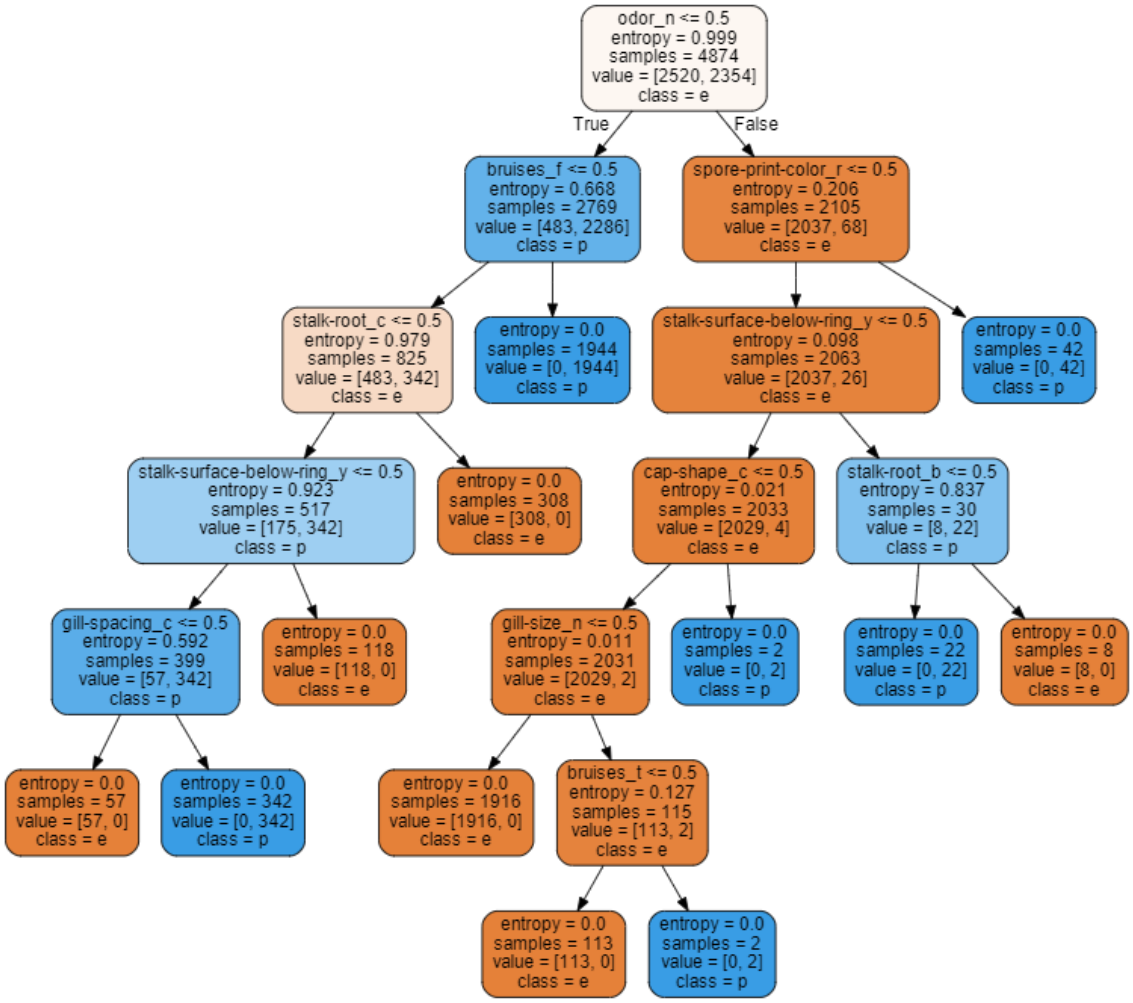
There are a total of 16 subsets of different proportions. (train/test) is converted to (train/total), so there are 4 proportions: 0.4, 0.6, 0.8 and 0.9. Each proportion splits the data sets into 4 subsets: X_train, X_test, y_train and y_test. All attributes were divided into n columns based on the unique values of that columns (format: name_value, Ex: odor_n, odor_p, etc.) and the values of that divided column are 0 for false and 1 for true (Ex: odor_n = 0 is equal to odor != n).

II. BUILDING THE DECISION TREE CLASSIFIERS:

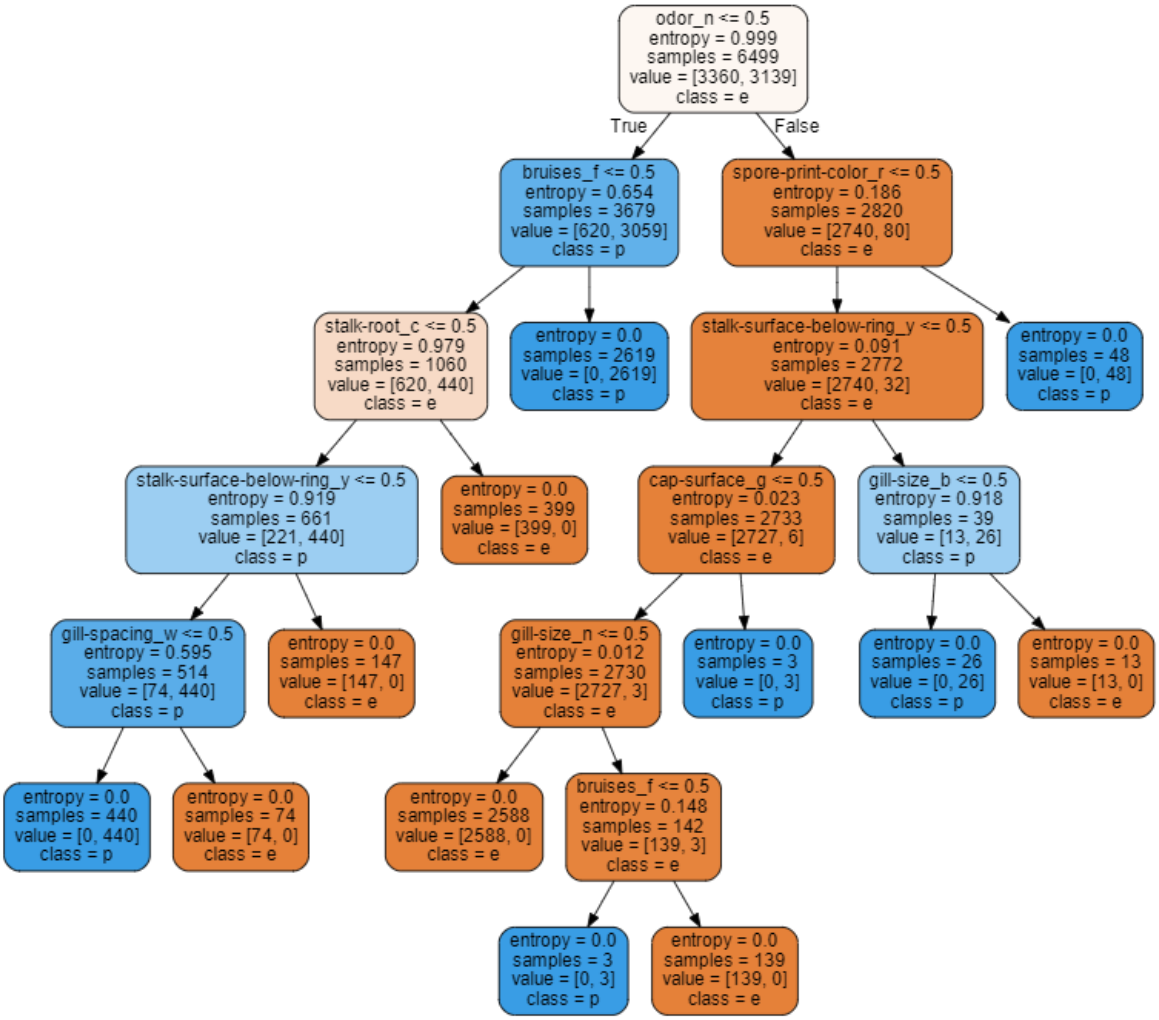
There are 4 decision tree classifiers needed to be built. The images below are from my experiment.



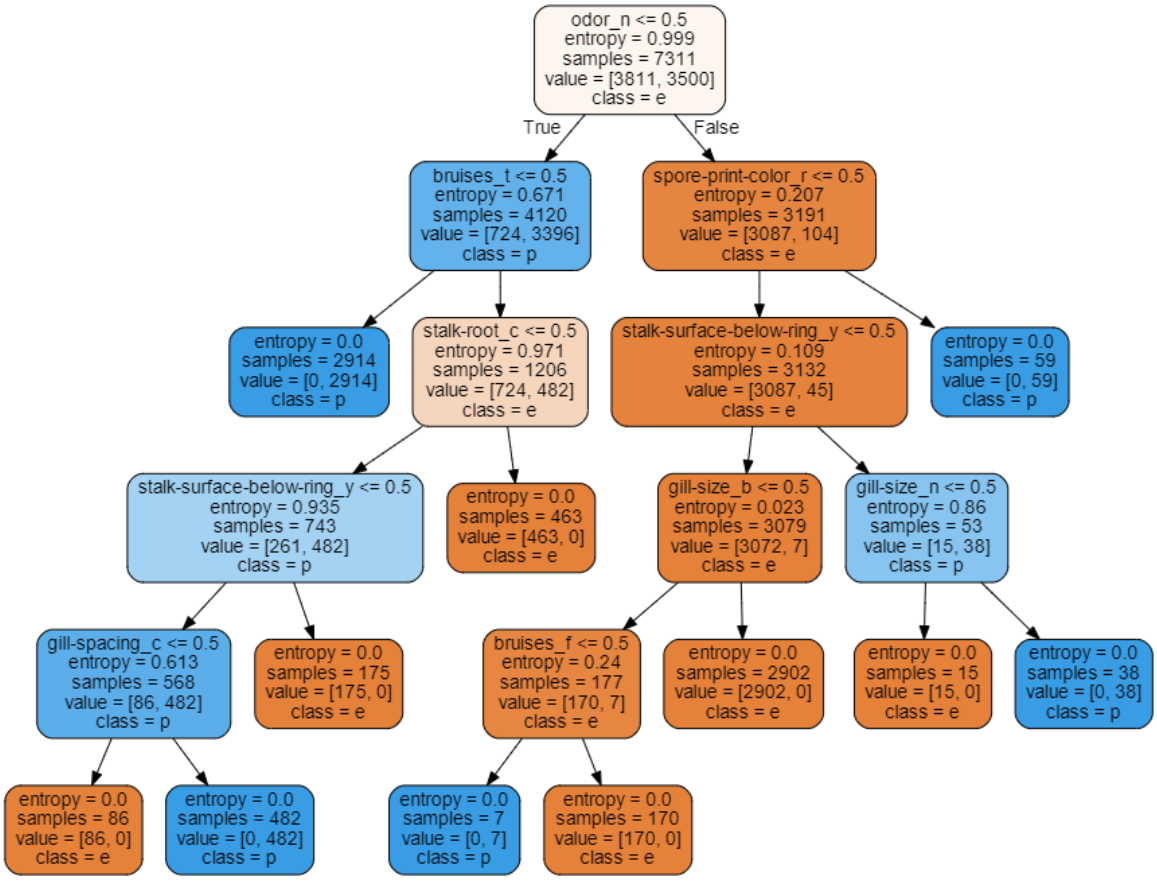
Proportion: 0.4 (train/test = 40/60) – ID: 1



Proportion: 0.6 (train/test = 60/40) – ID: 3



Proportion: 0.8 (train/test = 80/20) – ID: 2



Proportion: 0.9 (train/test = 90/10) – ID: 4

III. EVALUATING THE DECISION TREE CLASSIFIERS:

- Decision tree classifier #1:

Classification report

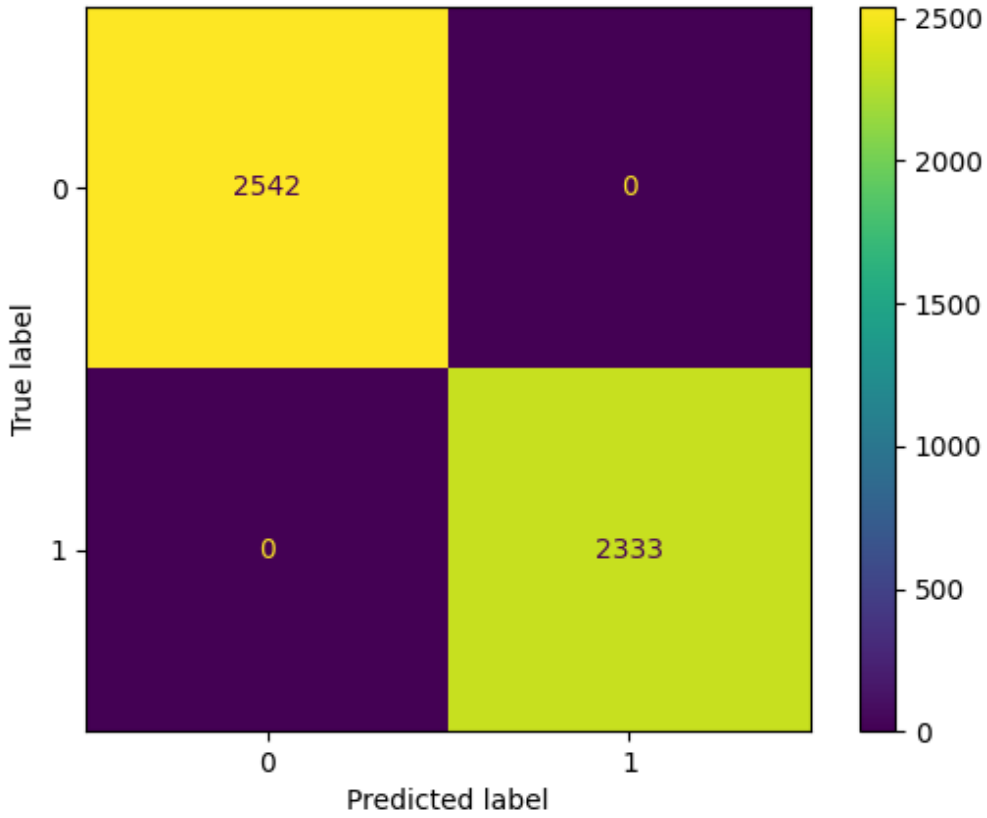
```
Report #1:
              precision    recall  f1-score   support

           e       1.00      1.00      1.00      2542
           p       1.00      1.00      1.00      2333

    accuracy                           1.00      4875
   macro avg       1.00      1.00      1.00      4875
weighted avg       1.00      1.00      1.00      4875

Confusion matrix was saved as output/confusion_matrix_1.png
```

Confusion matrix



- **Decision tree classifier #2:**

Classification report

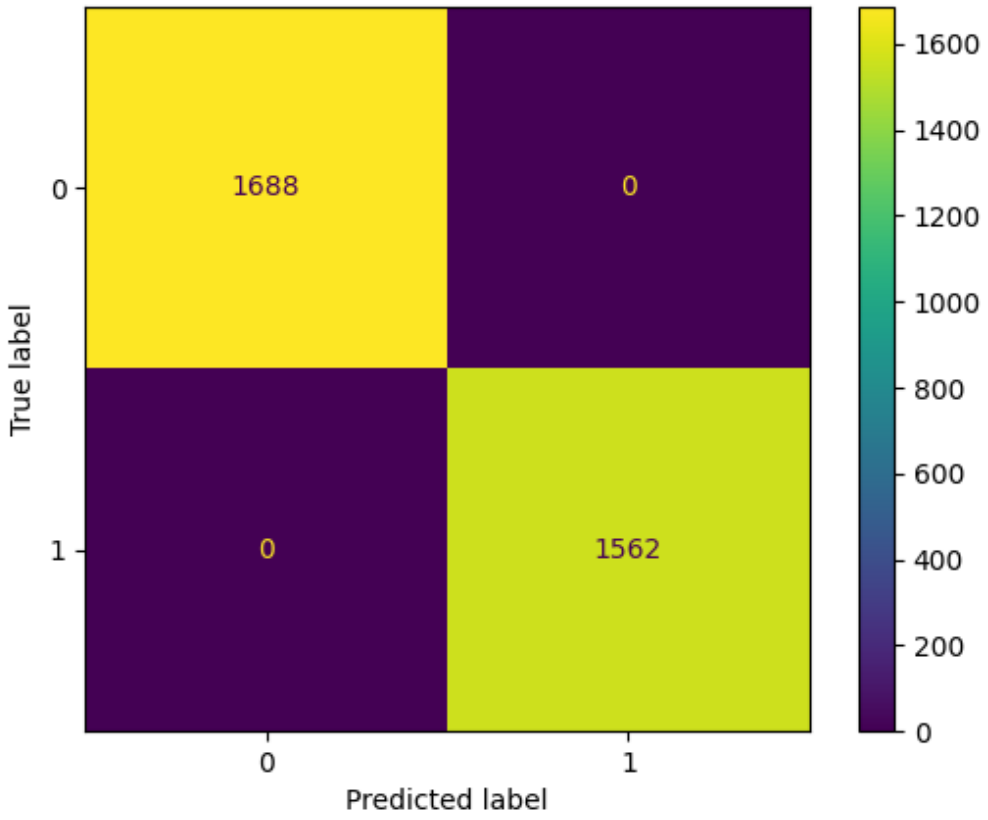
```
Report #2:
              precision    recall  f1-score   support

           e       1.00      1.00      1.00      1688
           p       1.00      1.00      1.00      1562

    accuracy                           1.00      3250
   macro avg       1.00      1.00      1.00      3250
weighted avg       1.00      1.00      1.00      3250

Confusion matrix was saved as output/confusion_matrix_2.png
```

Confusion matrix



- **Decision tree classifier #3:**

Classification report

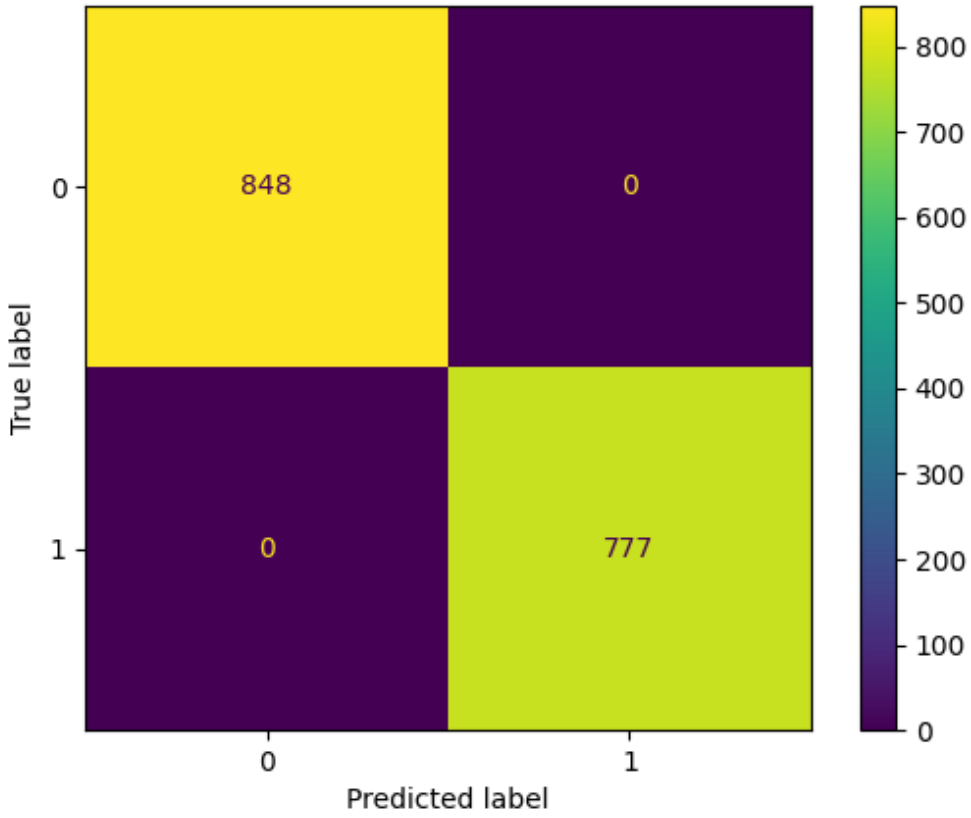
```
Report #3:
              precision    recall  f1-score   support

           e       1.00      1.00      1.00       848
           p       1.00      1.00      1.00       777

    accuracy                           1.00      1625
   macro avg       1.00      1.00      1.00      1625
weighted avg       1.00      1.00      1.00      1625

Confusion matrix was saved as output/confusion_matrix_3.png
```

Confusion matrix



- **Decision tree classifier #4:**

Classification report

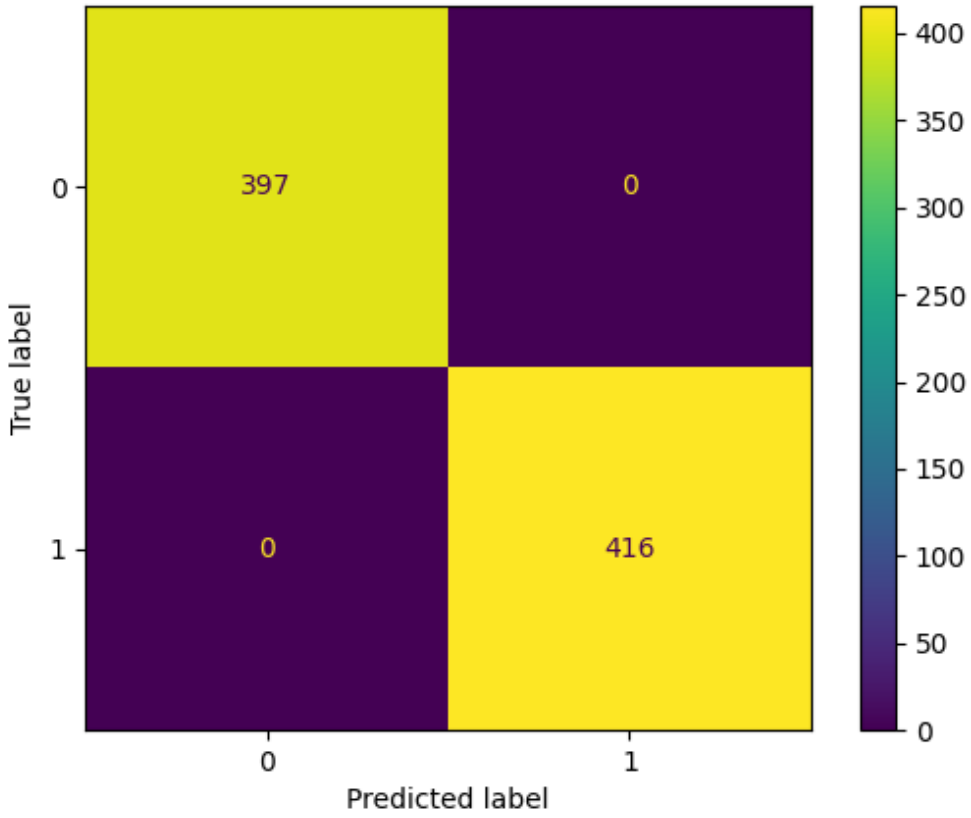
```
Report #4:
              precision    recall  f1-score   support

           e       1.00      1.00      1.00       397
           p       1.00      1.00      1.00       416

    accuracy                           1.00       813
   macro avg       1.00      1.00      1.00       813
weighted avg       1.00      1.00      1.00       813

Confusion matrix was saved as output/confusion_matrix_4.png
```

Confusion matrix



- # Comment:

Classification report explanation:

- Precision: What percent of predictions were correct? It is calculated by the ratio of true positives (the prediction and the case are the same, positive) to the sum of true and false positives (the prediction is positive but the case negative).
- Recall: What percent of the positive cases were caught? It is calculated by the ratio of true positives (the prediction and the case are the same, positive) to the sum of true and false negatives (the prediction is negative but the case positive).
- F1-score: What percent of positive predictions were correct? It is calculated based on Precision and Recall.

Confusion matrix explanation:

- It is a summary of prediction results.
- The number of correct and incorrect predictions are summerized with count values and broken down by each class.

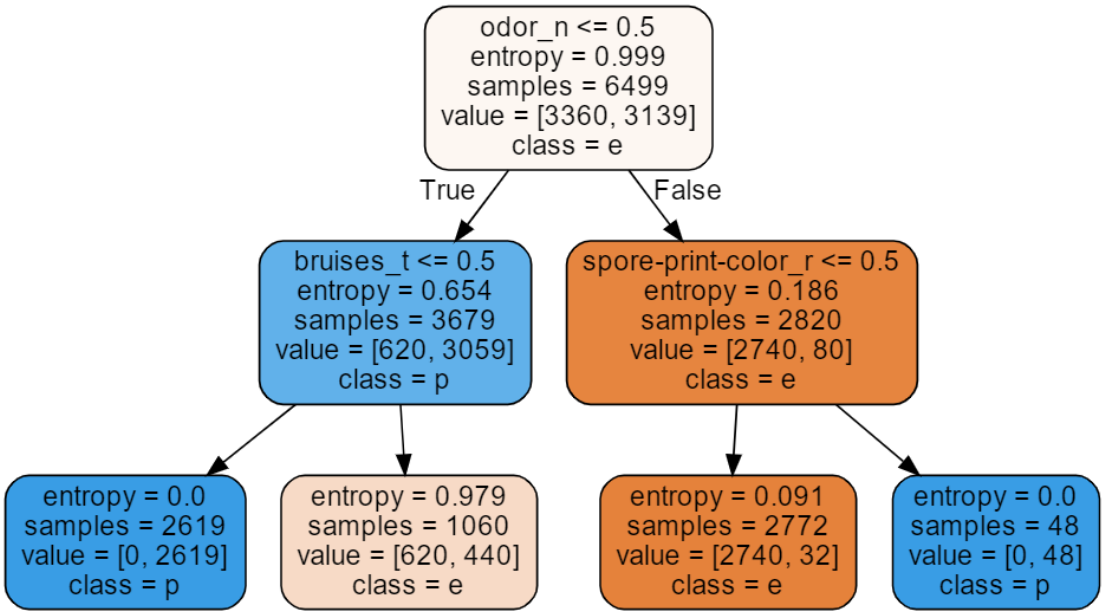
My comment:

- The performances of those decision tree classifiers are approximately the same.
- The accuracy is still enough even the proportion of train set is just 40% of the original data set.
- Sometimes there will be a few samples are predicted wrong, however it is not remarkable.

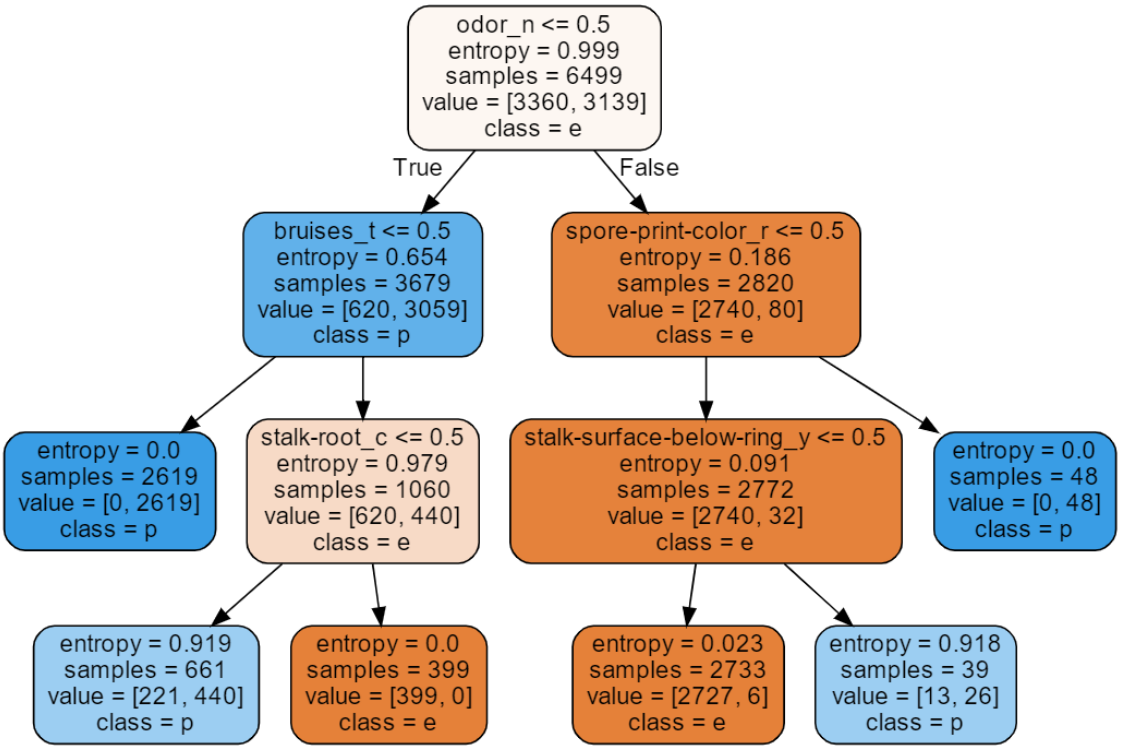
IV. THE DEPTH AND ACCURACY OF A DECISION TREE:

- **Decision tree classifier (depth = None):**

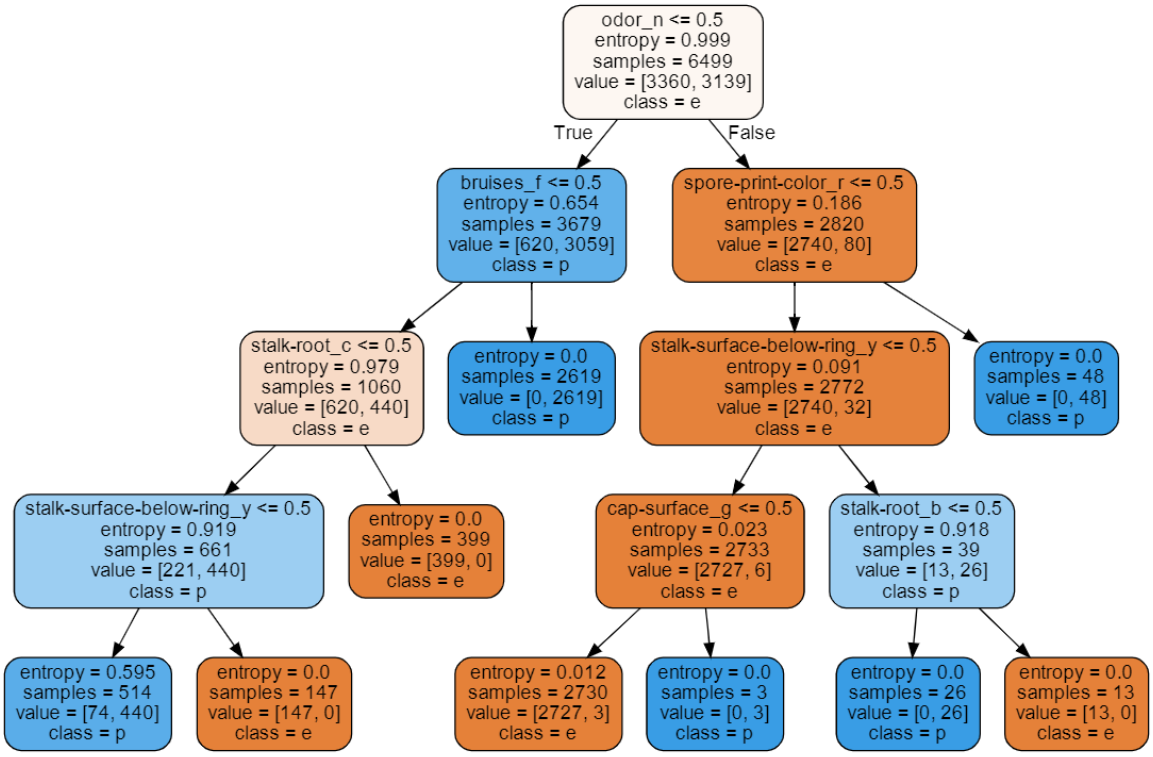
- **Decision tree classifier (depth = 2):**



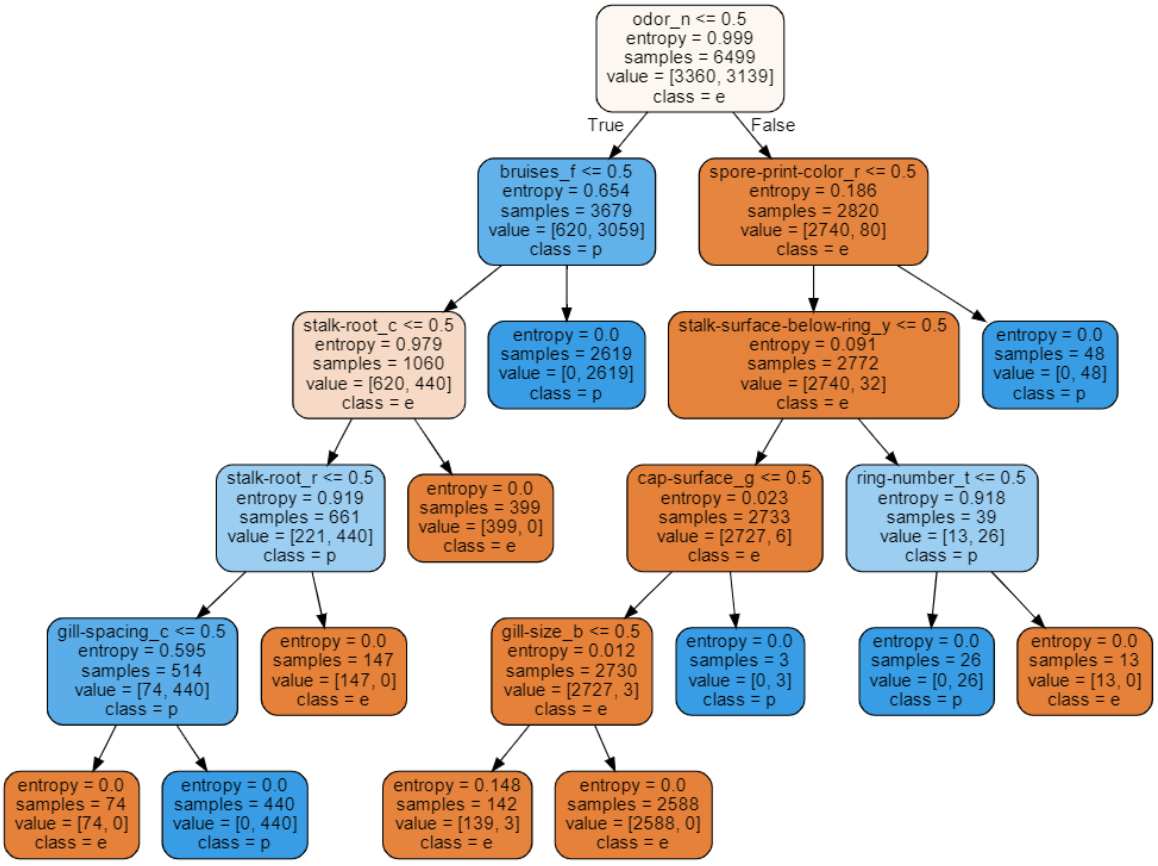
- **Decision tree classifier (depth = 3):**



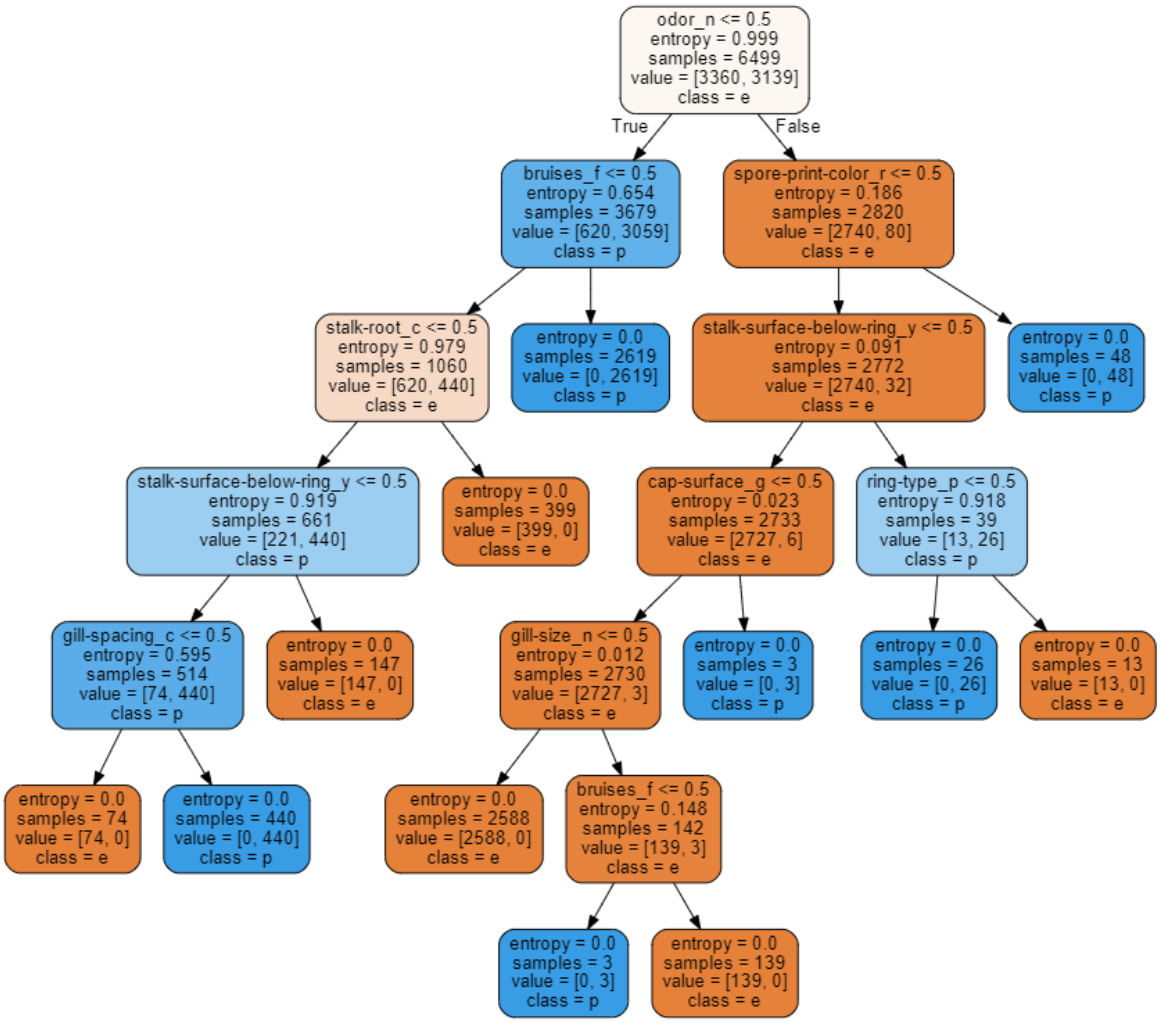
- **Decision tree classifier (depth = 4):**



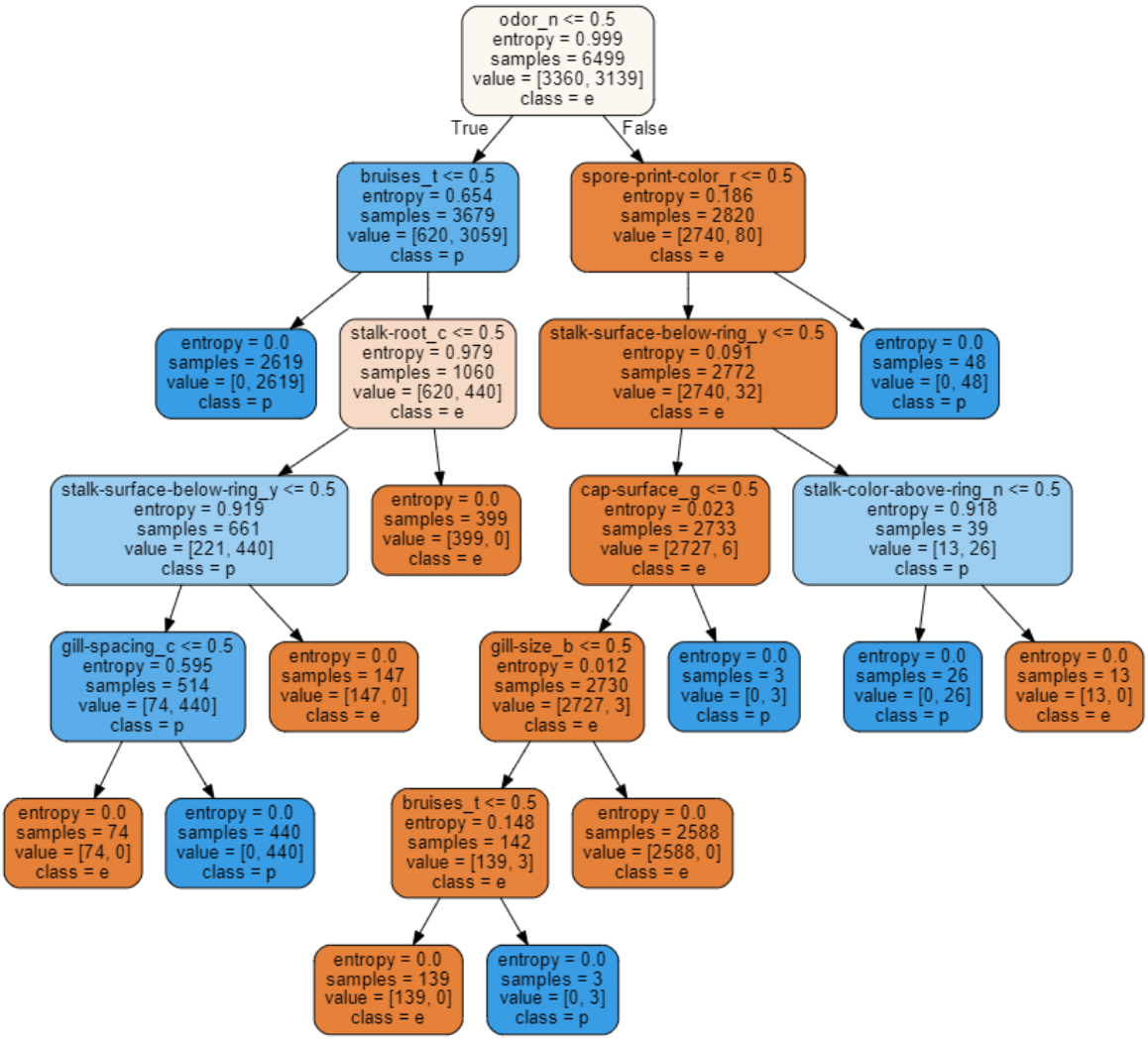
- **Decision tree classifier (depth = 5):**



- **Decision tree classifier (depth = 6):**



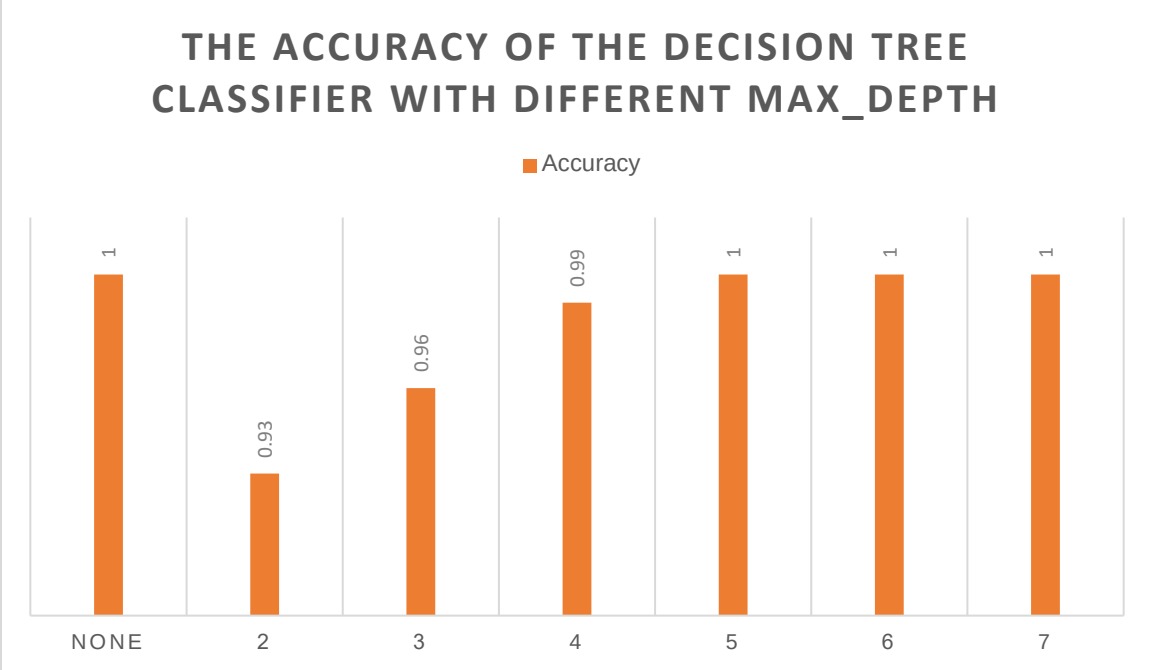
- **Decision tree classifier (depth = 7):**



- **Statistics:**

The information below was collected from my own experiment with 7 decision trees above.

Max_depth	**None**	**2**	**3**	**4**	**5**	**6**	**7**
Accuracy	1.00	0.93	0.96	0.99	1.00	1.00	1.00



- **Comment:**

  - Max_depth = None indicates that the tree was built exactly the same with the tree of proportion 0.8 in the building the decision tree classifiers section.
  - Max_depth increases, the accuracy also increases, correspondingly.
  - Max_depth = 2 is the least accurate among the others because it is insufficient depth to classify accurately.
  - At Max_depth = 5 and deeper, the depth is enough to predict correctly with this set.

V. REFERENCES:

1. https://youtu.be/q90UDEgYqeI
2. https://www.kaggle.com/haimfeld87/analysis-and-classification-of-mushrooms/notebook
3. https://scikit-learn.org/stable/index.html
4. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html
5. https://muthu.co/understanding-the-classification-report-in-sklearn/
6. https://machinelearningmastery.com/confusion-matrix-machine-learning/