Vietnam National University - Ho Chi Minh

University of Science
Faculty of Information Technology
.....o0o.....

LAB REPORT 02:
DECISION TREE WITH SCIKIT-LEARN

Course: Artificial Intelligence

Theory teacher:	**Nguyen Ngoc Thao**	
Instructors practice:	**Nguyen Thanh An**	
	Ho Thi Thanh Tuyen	
	Le Ngoc Thanh	
Student:	**Hoang Huu Minh An**	**ID: 20127102**
Class:	**20CLC01**	

Contents:

I. Information: 2
II. Introduction: 2
III. Assignment requirements: 2
a) Preparing the data sets: 2
b) Building the decision tree classifiers: 2
c) Evaluating the decision tree classifiers: 5
d) The depth and accuracy of a decision tree: 10
IV. List of References: 14

I. Information:

This Lab02: Decision tree presented by Hoang Huu Minh An-20127102.

II. Introduction:

About the Mushroom data set:

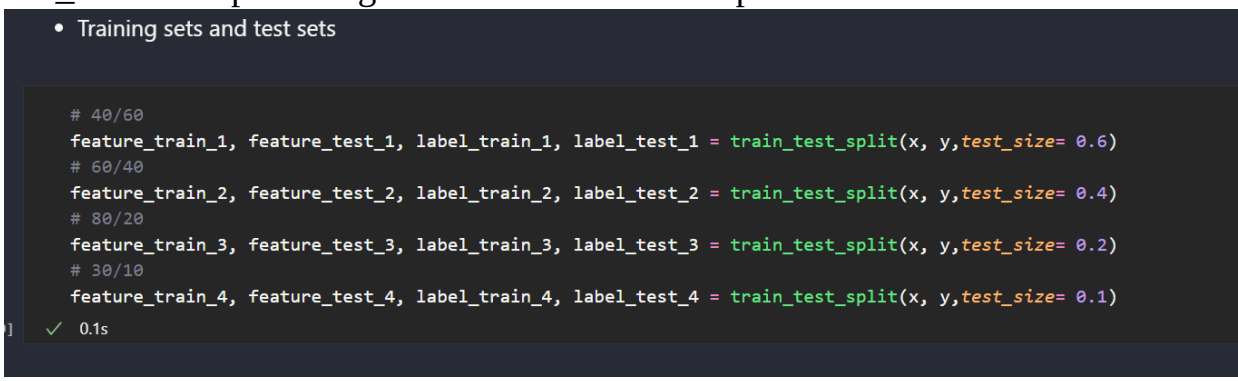
This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms. Each species is identified as either edible or poisonous. There are 8124 samples, each of which is characterized by 22 attributes and a target attribute.

- The target attribute is located at the first column
- The original data comes from the UCI Machine learning repository. You may refer to this page for more information

III. Assignment requirements:

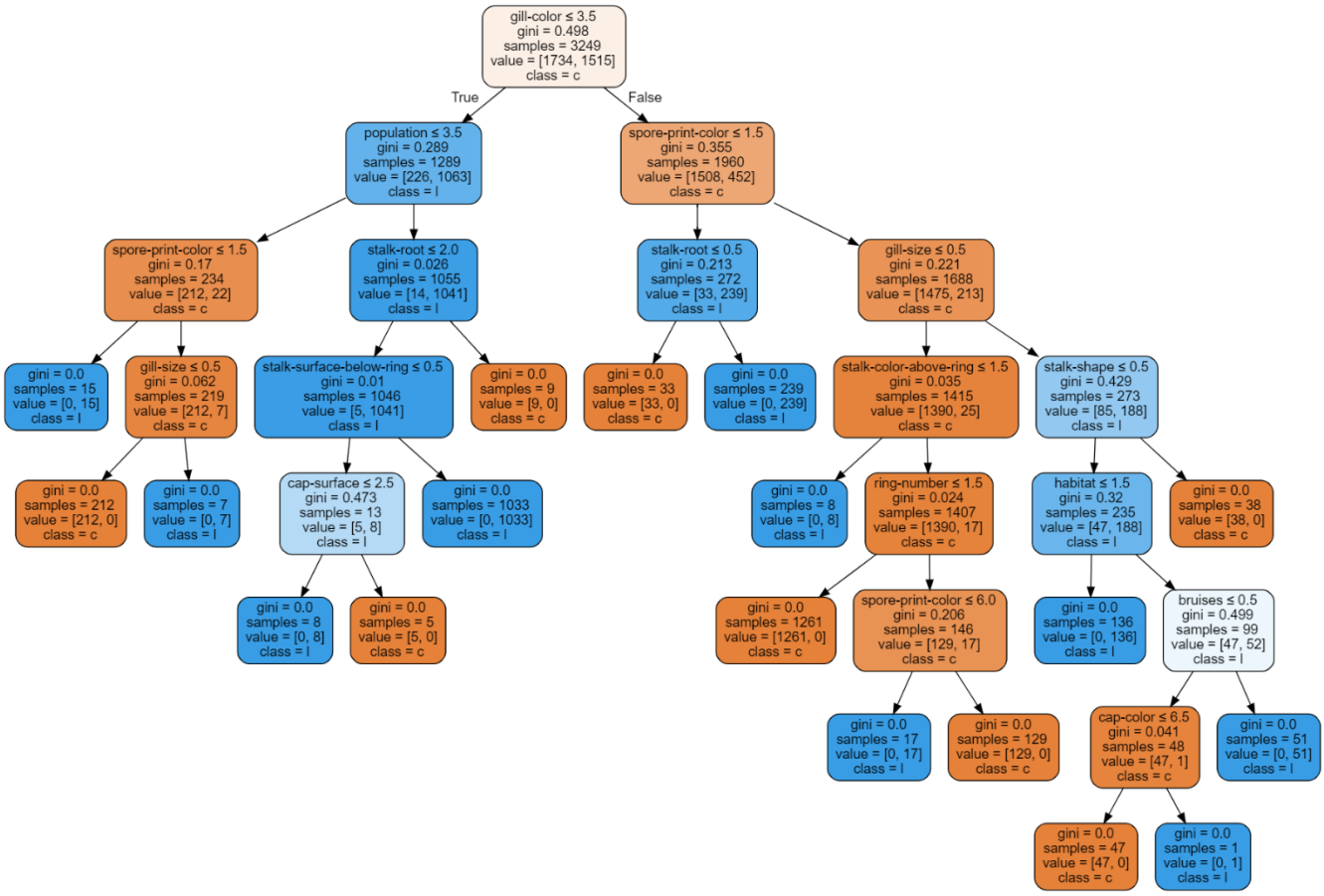
a) Preparing the data sets:

- Read data from file 'mushrooms.csv' using read_csv . function.
- Count the number of output ouputs
- Visualize columns of data
- Representing character data into corresponding representation data
- From the table above it can be seen that the column "veil-type" is 0 and not contributing to the data so I remove it.
- Shuffle data using the function shuffle
- Split the given data into the first column is output, the remaining 21 columns are data to train
- Split the data into train test sets using the split_split_test function with the x, y parameters respectively the data that I have split before, test_size is the percentage of data that I want to split .

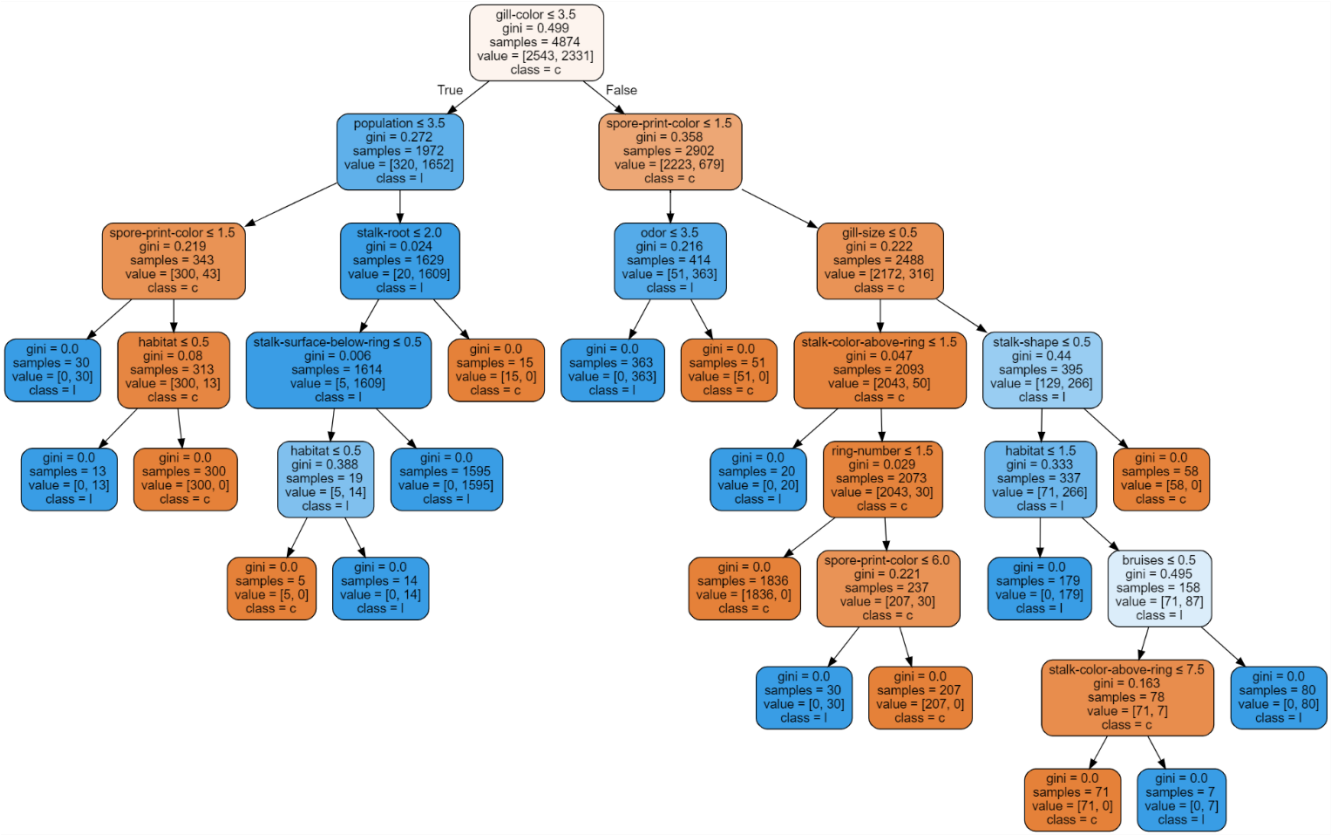


b) Building the decision tree classifiers:

➢ 40/60:

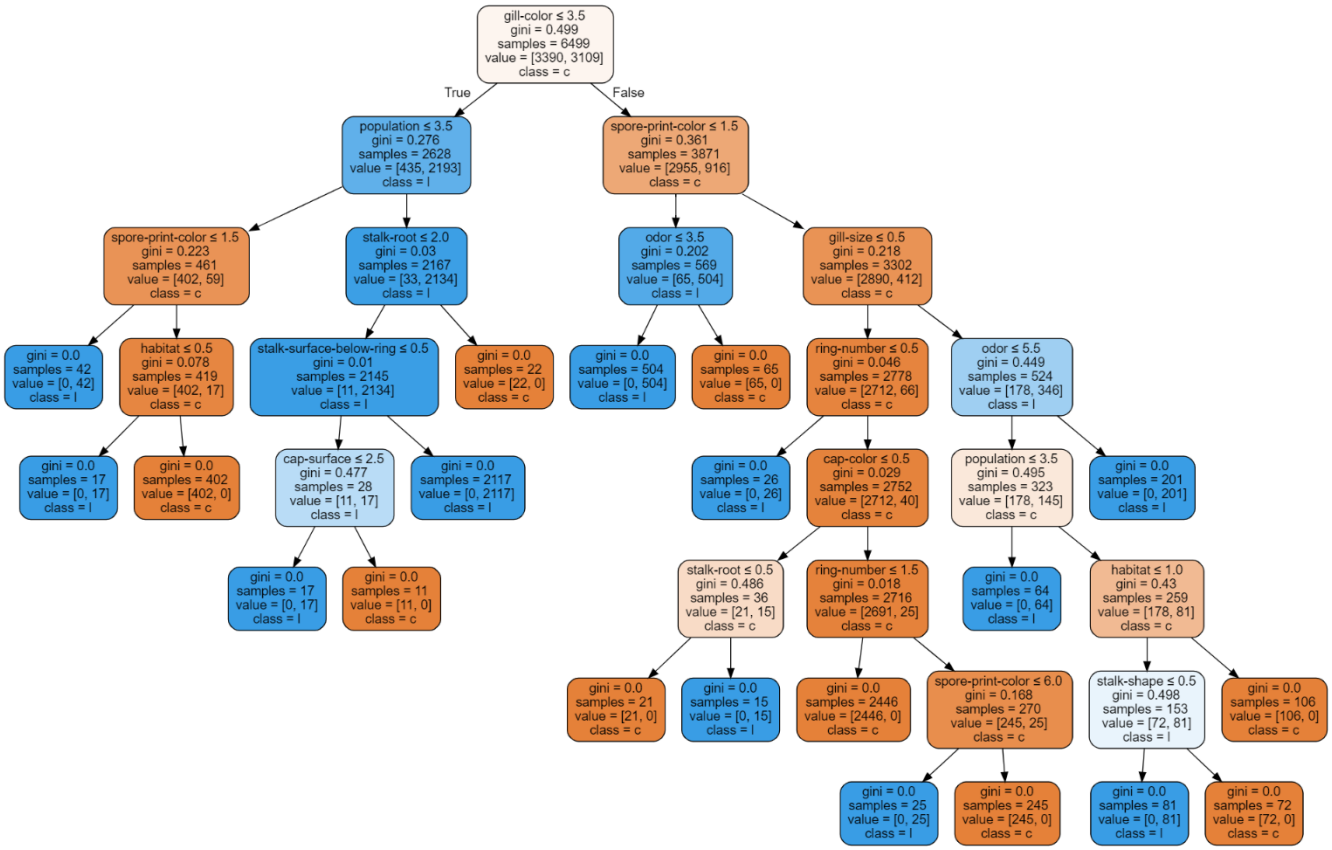


- 60/40

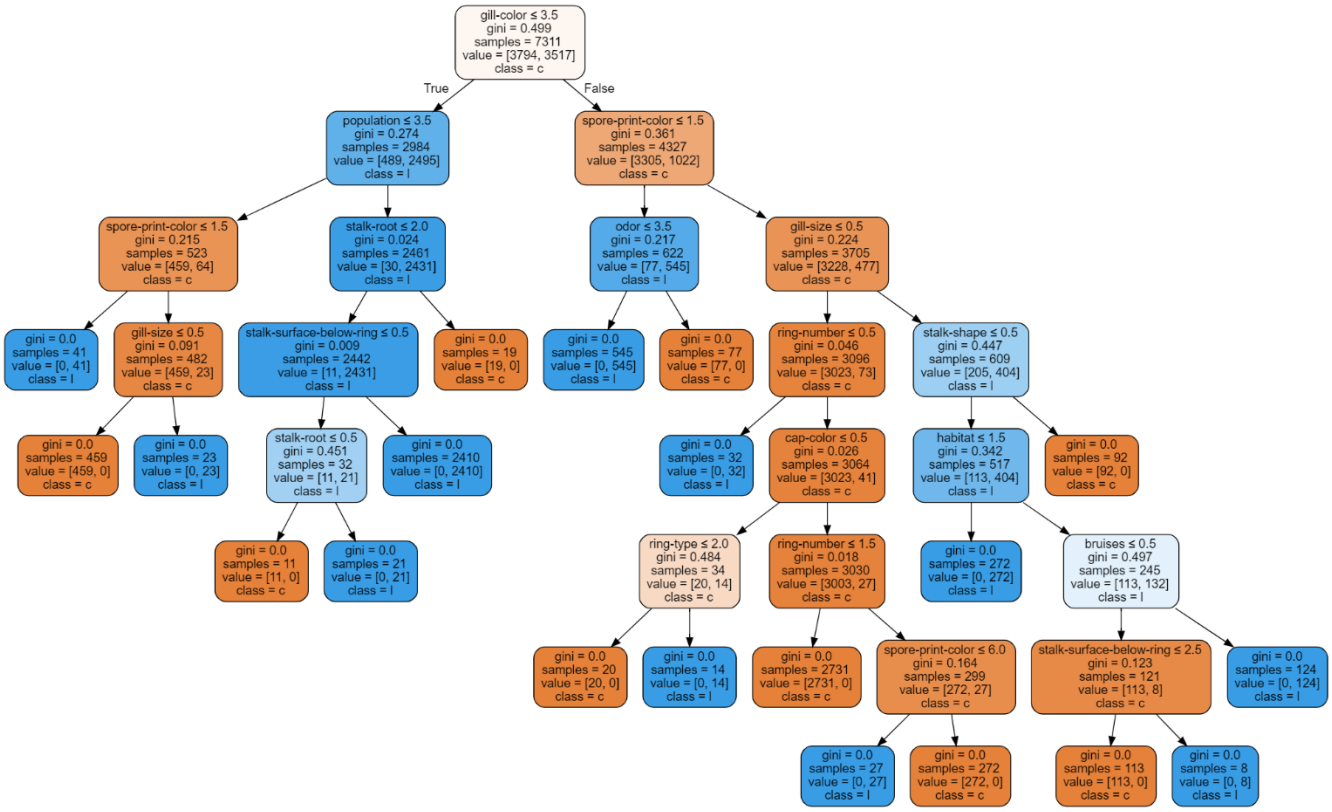


Hoang Huu Minh An - 20127102

- 80/20:


gill-color ≤ 3.5
gini = 0.499
samples = 6499
value = [3390, 3109]
class = c
True
False
population ≤ 3.5
gini = 0.276
samples = 2628
value = [435, 2193]
class = I
spore-print-color ≤ 1.5
gini = 0.361
samples = 3871
value = [2955, 916]
class = c
spore-print-color ≤ 1.5
gini = 0.223
samples = 461
value = [402, 59]
class = c
stalk-root ≤ 2.0
gini = 0.03
samples = 2167
value = [33, 2134]
class = I
odor ≤ 3.5
gini = 0.202
samples = 569
value = [65, 504]
class = I
gill-size ≤ 0.5
gini = 0.218
samples = 3302
value = [2890, 412]
class = c
gini = 0.0
samples = 42
value = [0, 42]
class = I
habitat ≤ 0.5
gini = 0.078
samples = 419
value = [402, 17]
class = c
stalk-surface-below-ring ≤ 0.5
gini = 0.01
samples = 2145
value = [11, 2134]
class = I
gini = 0.0
samples = 22
value = [22, 0]
class = c
gini = 0.0
samples = 504
value = [0, 504]
class = I
gini = 0.0
samples = 65
value = [65, 0]
class = c
ring-number ≤ 0.5
gini = 0.046
samples = 2778
value = [2712, 66]
class = c
odor ≤ 5.5
gini = 0.449
samples = 524
value = [178, 346]
class = I
gini = 0.0
samples = 17
value = [0, 17]
class = I
gini = 0.0
samples = 402
value = [402, 0]
class = c
cap-surface ≤ 2.5
gini = 0.477
samples = 28
value = [11, 17]
class = I
gini = 0.0
samples = 2117
value = [0, 2117]
class = I
gini = 0.0
samples = 26
value = [0, 26]
class = I
cap-color ≤ 0.5
gini = 0.029
samples = 2752
value = [2712, 40]
class = c
population ≤ 3.5
gini = 0.495
samples = 323
value = [178, 145]
class = c
gini = 0.0
samples = 201
value = [0, 201]
class = I
gini = 0.0
samples = 17
value = [0, 17]
class = I
gini = 0.0
samples = 11
value = [11, 0]
class = c
stalk-root ≤ 0.5
gini = 0.486
samples = 36
value = [21, 15]
class = c
ring-number ≤ 1.5
gini = 0.018
samples = 2716
value = [2691, 25]
class = c
gini = 0.0
samples = 64
value = [0, 64]
class = I
habitat ≤ 1.0
gini = 0.43
samples = 259
value = [178, 81]
class = c
gini = 0.0
samples = 21
value = [21, 0]
class = c
gini = 0.0
samples = 15
value = [0, 15]
class = I
gini = 0.0
samples = 2446
value = [2446, 0]
class = c
spore-print-color ≤ 6.0
gini = 0.168
samples = 270
value = [245, 25]
class = c
stalk-shape ≤ 0.5
gini = 0.498
samples = 153
value = [72, 81]
class = I
gini = 0.0
samples = 106
value = [106, 0]
class = c
gini = 0.0
samples = 25
value = [0, 25]
class = I
gini = 0.0
samples = 245
value = [245, 0]
class = c
gini = 0.0
samples = 81
value = [0, 81]
class = I
gini = 0.0
samples = 72
value = [72, 0]
class = c

➢ 90/10:

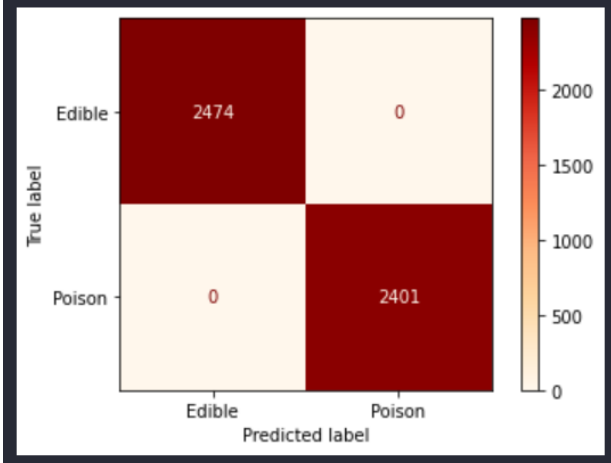


c) Evaluating the decision tree classifiers:

Classification tree predict a qualative response, in contrast to regression tree that predict quantitive response. Classification tree predicts the observation that belong to the most common occurring class from the training data for each region. In order to grow the classification tree the algorithm uses recursive binary splitting.

➢ 40/60:

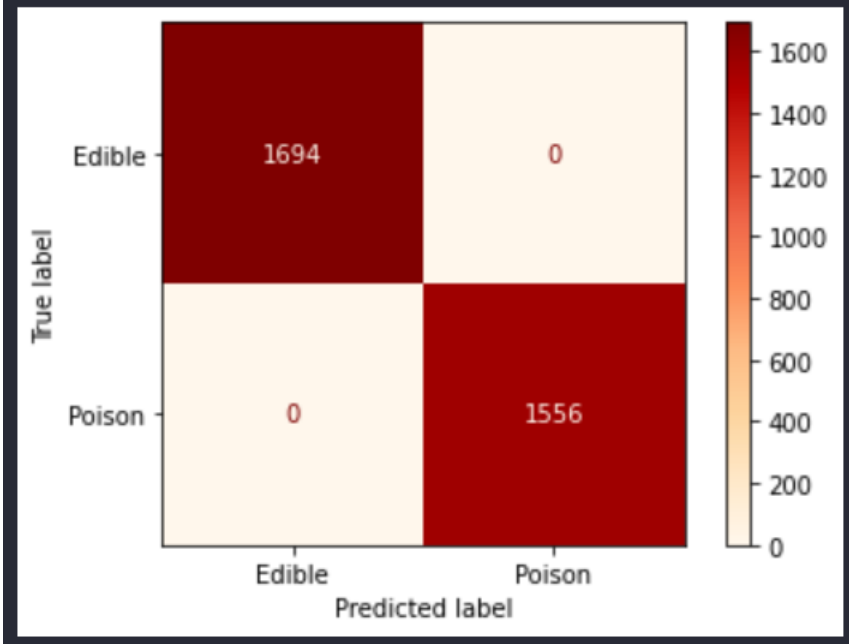
```
Accurary:  1.0
-----------------------------------------
Confusion matrix

c:\Program Files\Python310\lib\site-packages\sklearn\uti
`plot_confusion_matrix` is deprecated in 1.0 and will be
ConfusionMatrixDisplay.from_estimator.
  warnings.warn(msg, category=FutureWarning)
```




```
Decision Tree Classifier report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2474
           1       1.00      1.00      1.00      2401

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

-----------------------------------------------
AUC Score:  1.0
```

- 60/40:

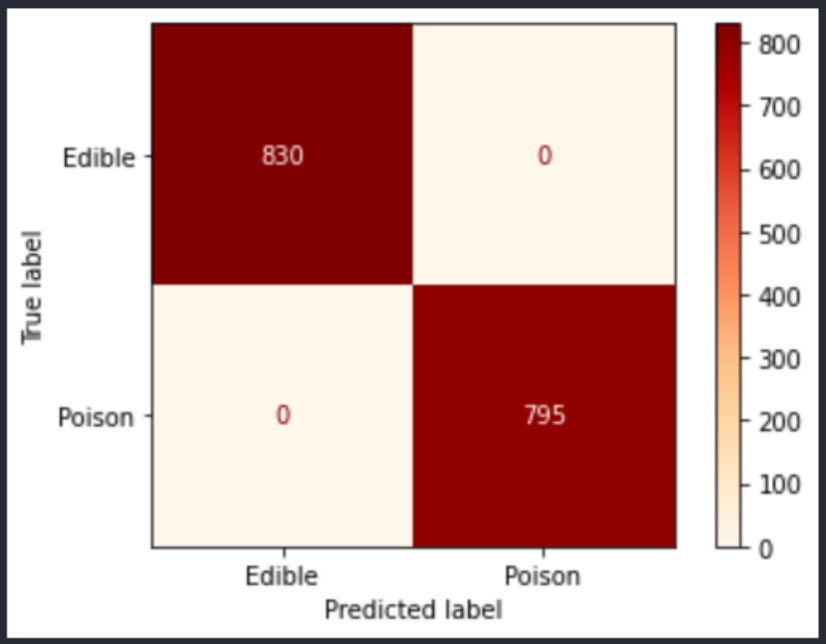
```
Accurary:  1.0
--------------------------------------------
Confusion matrix

c:\Program Files\Python310\lib\site-packages\sklea
`plot_confusion_matrix` is deprecated in 1.0 and w
ConfusionMatrixDisplay.from_estimator.
  warnings.warn(msg, category=FutureWarning)
```




```
Decision Tree Classifier report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1665
           1       1.00      1.00      1.00      1585

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

--------------------------------------------
AUC Score:  1.0
```

- 80/20:

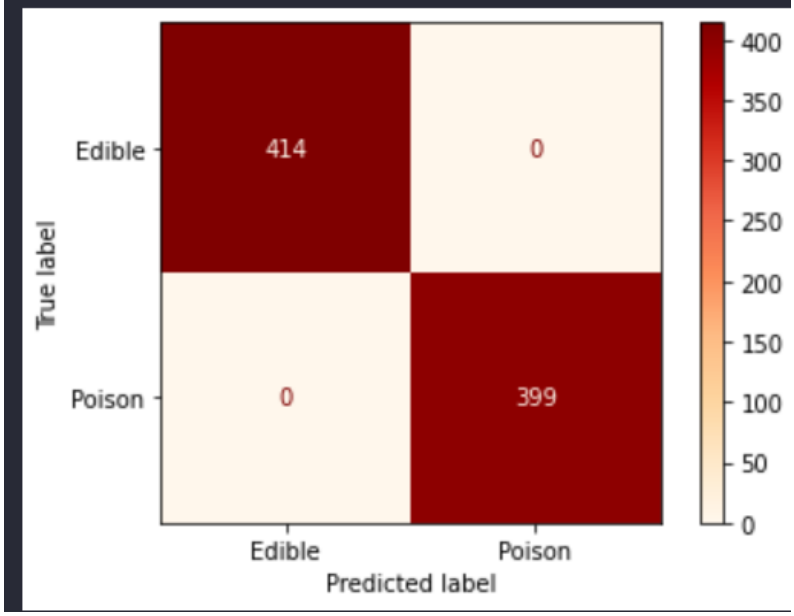
```
Accurary:  1.0
------------------------------------------------
Confusion matrix

c:\Program Files\Python310\lib\site-packages\sklear
`plot_confusion_matrix` is deprecated in 1.0 and wi
ConfusionMatrixDisplay.from_estimator.
  warnings.warn(msg, category=FutureWarning)
```



```
Decision Tree Classifier report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       818
           1       1.00      1.00      1.00       807

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

-----------------------------------------------
AUC Score:  1.0
```

- 90/10:

```
Accurary:  1.0
-----------------------------------------------
Confusion matrix

c:\Program Files\Python310\lib\site-packages
`plot_confusion_matrix` is deprecated in 1.0
ConfusionMatrixDisplay.from_estimator.
  warnings.warn(msg, category=FutureWarning)
```

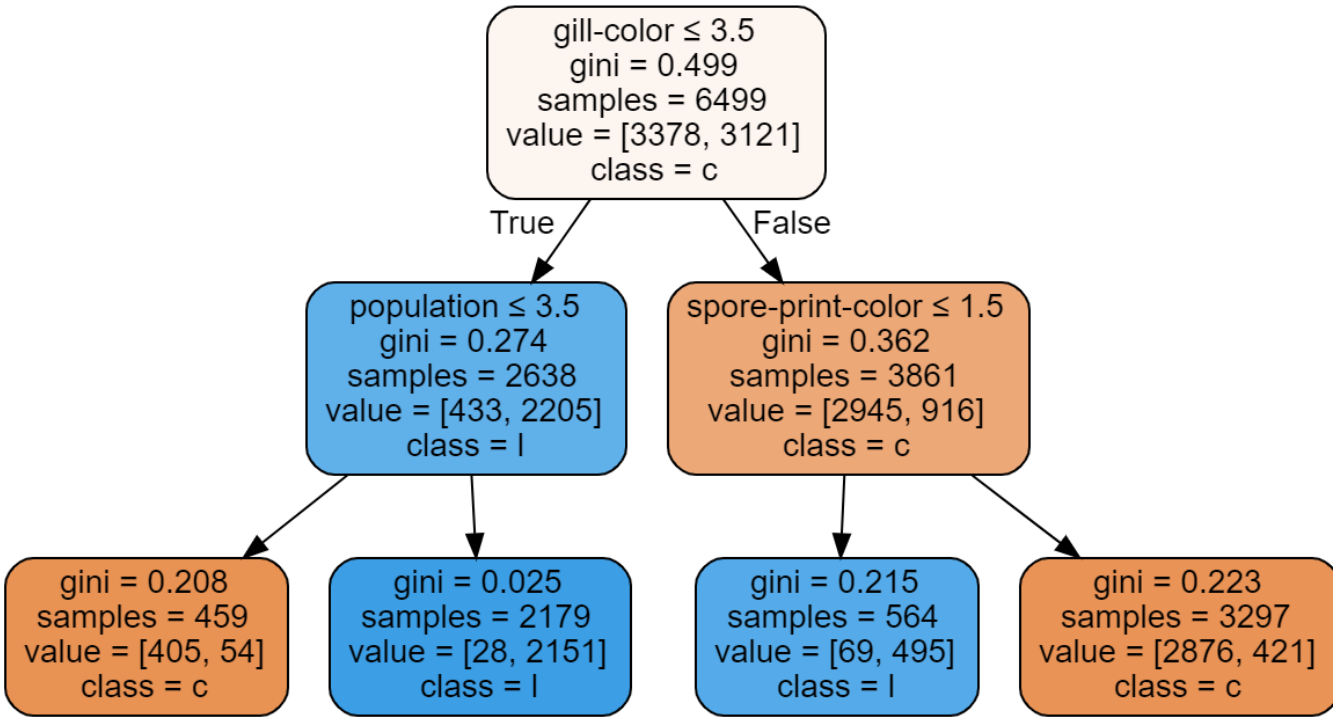


```
Decision Tree Classifier report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       414
           1       1.00      1.00      1.00       399

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

-----------------------------------------------
AUC Score:  1.0
```

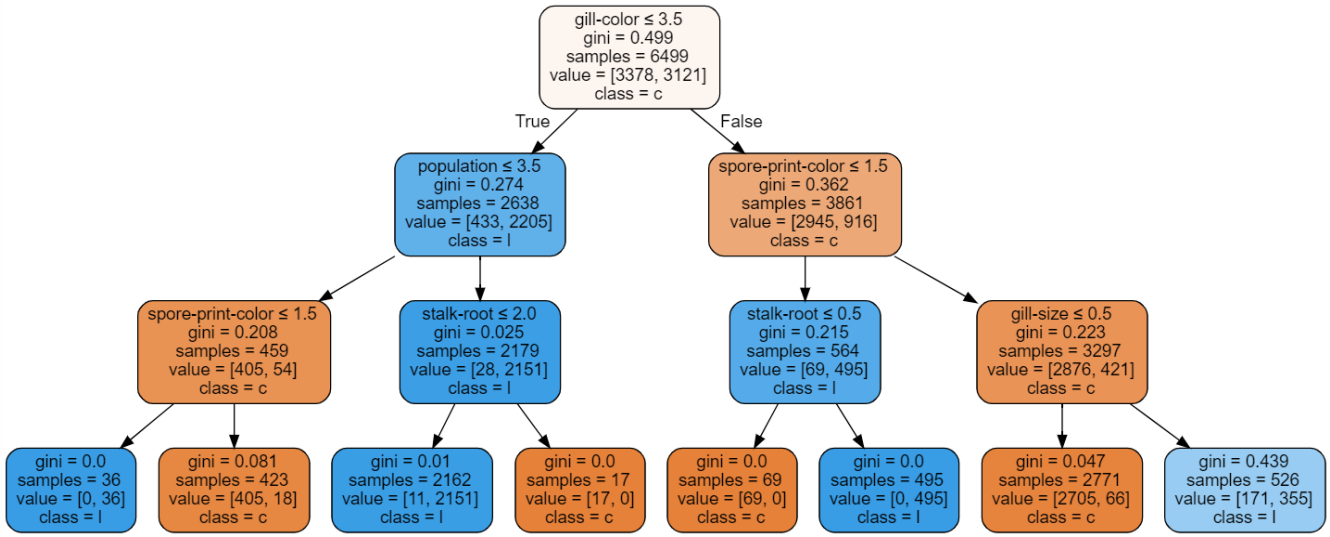
- They are some of the mostly used measures in evaluating how good your system works.
- **Precision:** It is calculated with respect to the predicted values. For class-A, out of total predictions how many were really belong to class-A in actual dataset, is defined as the precision. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] column.
- **Recall:** It is calculated with respect to the actual values in dataset. For class-A, out of total entries in dataset, how many were actually classified in class-A by the ML model, is defined as the recall. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] row.
- **F1-score:** It is the harmonic mean of precision and recall.
- **Support:** It is the total entries of each class in the actual dataset. It is simply the sum of rows for every class-i

d) The depth and accuracy of a decision tree:

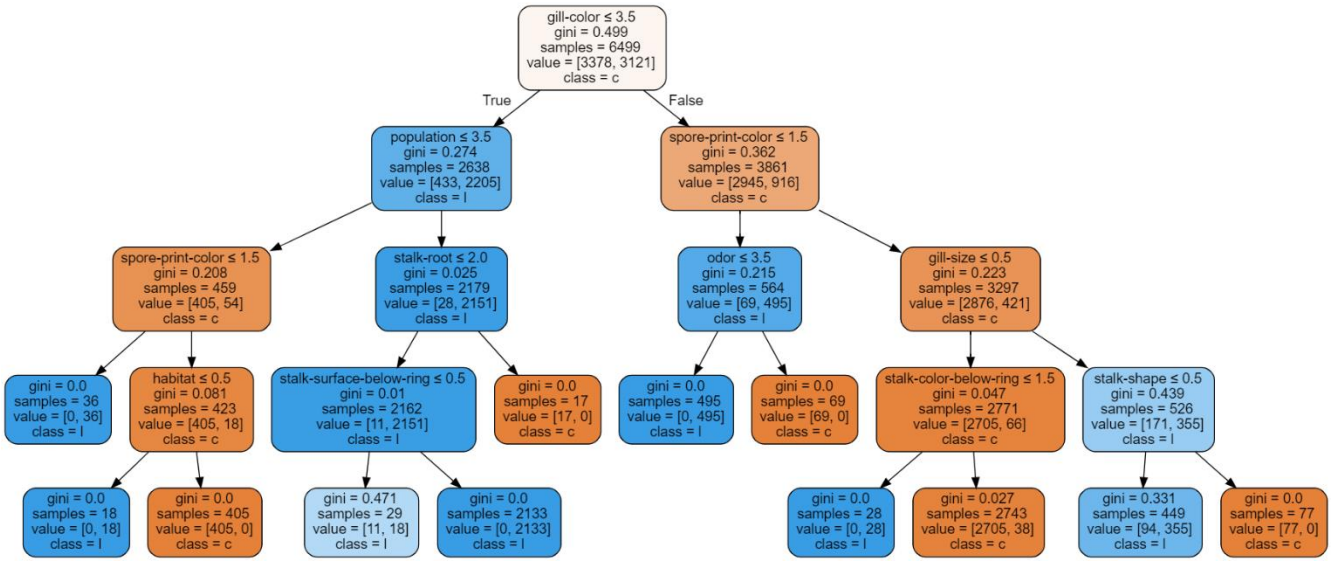
- **Max_depth: 2**



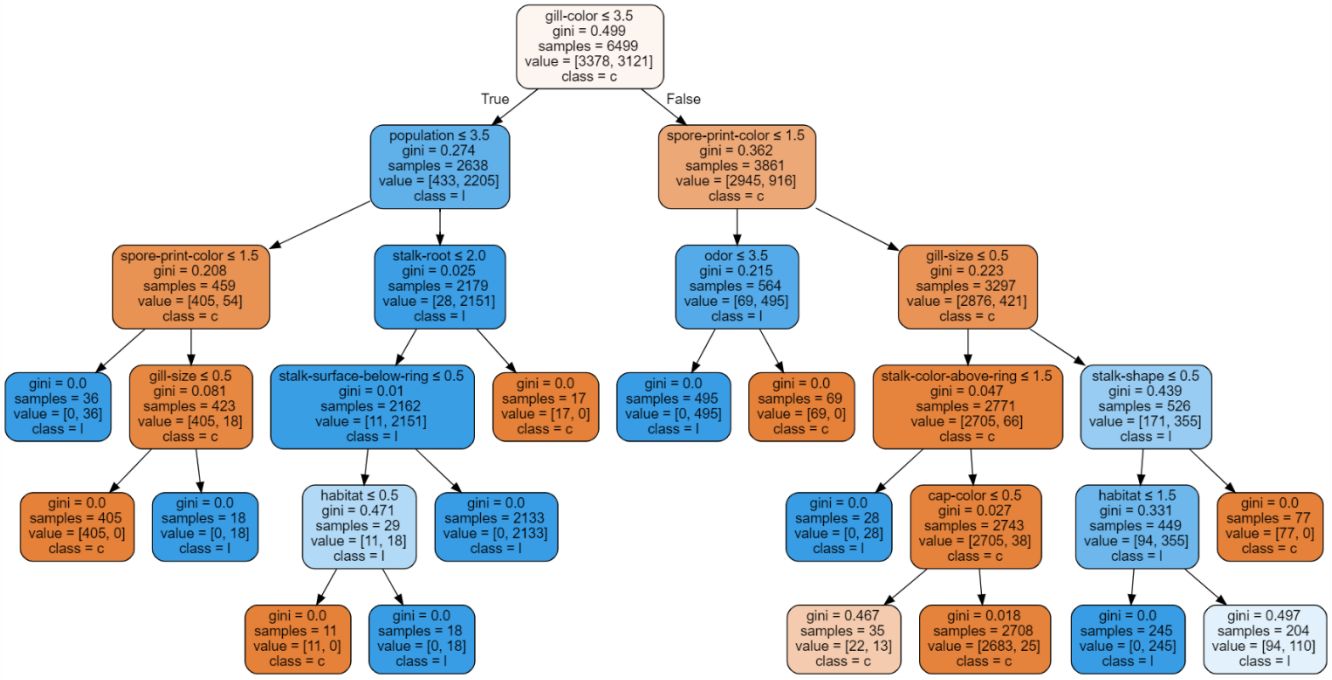
- **Max_depth: 3**



- **Max_depth: 4**



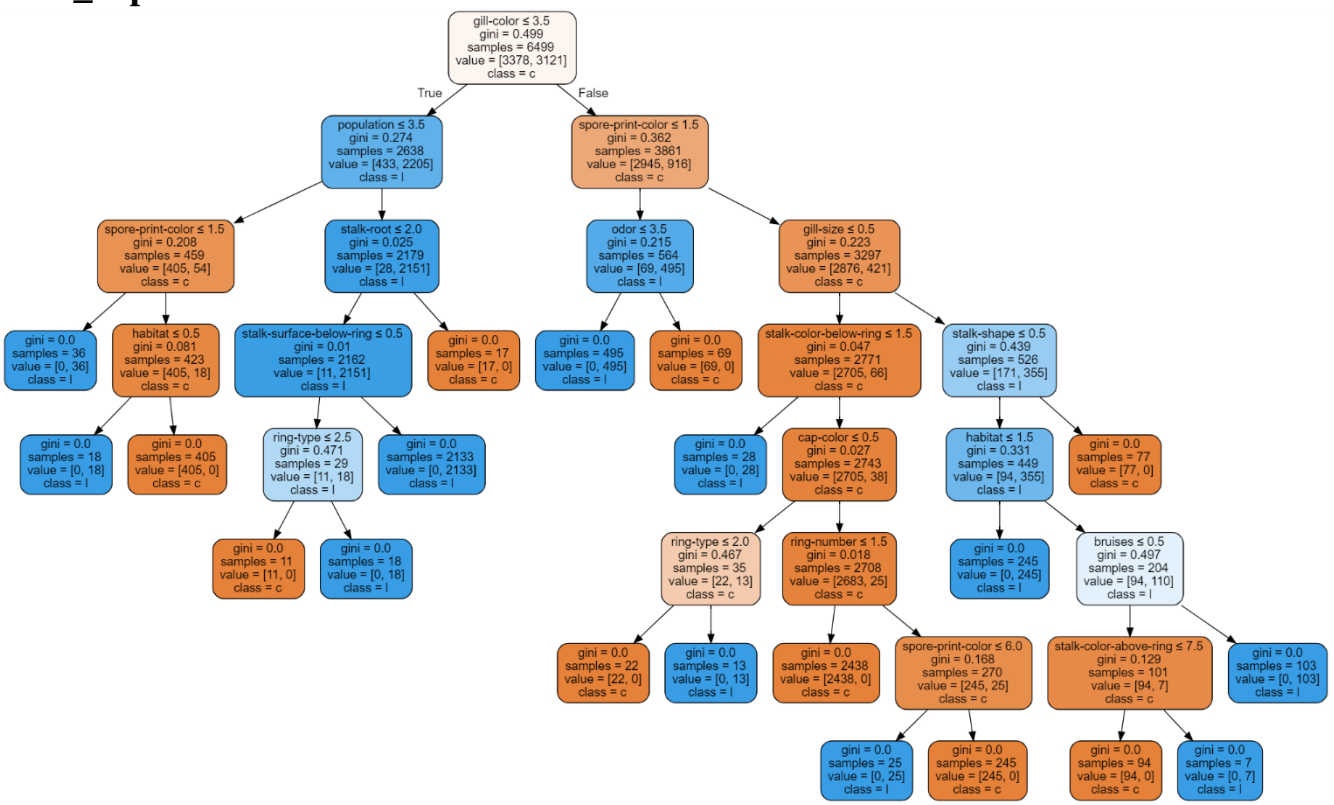
- **Max_depth: 5**



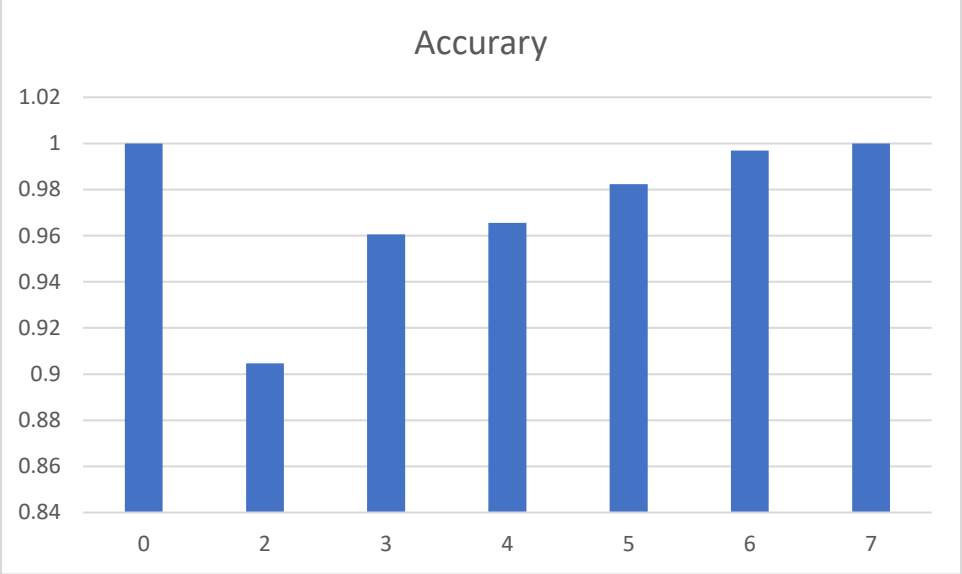
- **Max_depth: 6**


gill-color ≤ 3.5
gini = 0.499
samples = 6499
value = [3378, 3121]
class = c
True
False
population ≤ 3.5
gini = 0.274
samples = 2638
value = [433, 2205]
class = I
spore-print-color ≤ 1.5
gini = 0.362
samples = 3861
value = [2945, 916]
class = c
spore-print-color ≤ 1.5
gini = 0.208
samples = 459
value = [405, 54]
class = c
stalk-root ≤ 2.0
gini = 0.025
samples = 2179
value = [28, 2151]
class = I
odor ≤ 3.5
gini = 0.215
samples = 564
value = [69, 495]
class = I
gill-size ≤ 0.5
gini = 0.223
samples = 3297
value = [2876, 421]
class = c
gini = 0.0
samples = 36
value = [0, 36]
class = I
gill-size ≤ 0.5
gini = 0.081
samples = 423
value = [405, 18]
class = c
stalk-surface-below-ring ≤ 0.5
gini = 0.01
samples = 2162
value = [11, 2151]
class = I
gini = 0.0
samples = 17
value = [17, 0]
class = c
gini = 0.0
samples = 495
value = [0, 495]
class = I
gini = 0.0
samples = 69
value = [69, 0]
class = c
ring-number ≤ 0.5
gini = 0.047
samples = 2771
value = [2705, 66]
class = c
stalk-shape ≤ 0.5
gini = 0.439
samples = 526
value = [171, 355]
class = I
gini = 0.0
samples = 405
value = [405, 0]
class = c
gini = 0.0
samples = 18
value = [0, 18]
class = I
bruises ≤ 0.5
gini = 0.471
samples = 29
value = [11, 18]
class = I
gini = 0.0
samples = 2133
value = [0, 2133]
class = I
gini = 0.0
samples = 28
value = [0, 28]
class = I
cap-color ≤ 0.5
gini = 0.027
samples = 2743
value = [2705, 38]
class = c
habitat ≤ 1.5
gini = 0.331
samples = 449
value = [94, 355]
class = I
gini = 0.0
samples = 77
value = [77, 0]
class = c
gini = 0.0
samples = 11
value = [11, 0]
class = c
gini = 0.0
samples = 18
value = [0, 18]
class = I
stalk-root ≤ 0.5
gini = 0.467
samples = 35
value = [22, 13]
class = c
ring-number ≤ 1.5
gini = 0.018
samples = 2708
value = [2683, 25]
class = c
gini = 0.0
samples = 245
value = [0, 245]
class = I
bruises ≤ 0.5
gini = 0.497
samples = 204
value = [94, 110]
class = I
gini = 0.0
samples = 22
value = [22, 0]
class = c
gini = 0.0
samples = 13
value = [0, 13]
class = I
gini = 0.0
samples = 2438
value = [2438, 0]
class = c
gini = 0.168
samples = 270
value = [245, 25]
class = c
gini = 0.129
samples = 101
value = [94, 7]
class = c
gini = 0.0
samples = 103
value = [0, 103]
class = I

- **Max_depth: 7**


gill-color ≤ 3.5
gini = 0.499
samples = 6499
value = [3378, 3121]
class = c
True
False
population ≤ 3.5
gini = 0.274
samples = 2638
value = [433, 2205]
class = I
spore-print-color ≤ 1.5
gini = 0.362
samples = 3861
value = [2945, 916]
class = c
spore-print-color ≤ 1.5
gini = 0.208
samples = 459
value = [405, 54]
class = c
stalk-root ≤ 2.0
gini = 0.025
samples = 2179
value = [28, 2151]
class = I
odor ≤ 3.5
gini = 0.215
samples = 564
value = [69, 495]
class = I
gill-size ≤ 0.5
gini = 0.223
samples = 3297
value = [2876, 421]
class = c
gini = 0.0
samples = 36
value = [0, 36]
class = I
habitat ≤ 0.5
gini = 0.081
samples = 423
value = [405, 18]
class = c
stalk-surface-below-ring ≤ 0.5
gini = 0.01
samples = 2162
value = [11, 2151]
class = I
gini = 0.0
samples = 17
value = [17, 0]
class = c
gini = 0.0
samples = 495
value = [0, 495]
class = I
gini = 0.0
samples = 69
value = [69, 0]
class = c
stalk-color-below-ring ≤ 1.5
gini = 0.047
samples = 2771
value = [2705, 66]
class = c
stalk-shape ≤ 0.5
gini = 0.439
samples = 526
value = [171, 355]
class = I
gini = 0.0
samples = 18
value = [0, 18]
class = I
gini = 0.0
samples = 405
value = [405, 0]
class = c
ring-type ≤ 2.5
gini = 0.471
samples = 29
value = [11, 18]
class = I
gini = 0.0
samples = 2133
value = [0, 2133]
class = I
gini = 0.0
samples = 28
value = [0, 28]
class = I
cap-color ≤ 0.5
gini = 0.027
samples = 2743
value = [2705, 38]
class = c
habitat ≤ 1.5
gini = 0.331
samples = 449
value = [94, 355]
class = I
gini = 0.0
samples = 77
value = [77, 0]
class = c
gini = 0.0
samples = 11
value = [11, 0]
class = c
gini = 0.0
samples = 18
value = [0, 18]
class = I
ring-type ≤ 2.0
gini = 0.467
samples = 35
value = [22, 13]
class = c
ring-number ≤ 1.5
gini = 0.018
samples = 2708
value = [2683, 25]
class = c
gini = 0.0
samples = 245
value = [0, 245]
class = I
bruises ≤ 0.5
gini = 0.497
samples = 204
value = [94, 110]
class = I
gini = 0.0
samples = 22
value = [22, 0]
class = c
gini = 0.0
samples = 13
value = [0, 13]
class = I
gini = 0.0
samples = 2438
value = [2438, 0]
class = c
spore-print-color ≤ 6.0
gini = 0.168
samples = 270
value = [245, 25]
class = c
stalk-color-above-ring ≤ 7.5
gini = 0.129
samples = 101
value = [94, 7]
class = c
gini = 0.0
samples = 103
value = [0, 103]
class = I
gini = 0.0
samples = 25
value = [0, 25]
class = I
gini = 0.0
samples = 245
value = [245, 0]
class = c
gini = 0.0
samples = 94
value = [94, 0]
class = c
gini = 0.0
samples = 7
value = [0, 7]
class = I

Max_deapth	None	2	3	4	5	6	7
Accurary	1.0	0.9046	0.9606	0.9655	0.9823	0.9969	1



- Length max tell me how deep can the tree be so when there is one node when there is one decision that's depth ones so when you make the decision or when there's only one node then that's a depth of one when you have a decision node there and then you split. Sometimes decision trees can grow to be very long so if you wanted something these things are not really things that you can.

IV. List of References:

[1] Scikit-learn decision trees: https://scikit-learn.org/stable/modules/tree.html

[2] Analysis and classification of Mushrooms: https://www.kaggle.com/haimfeld87/analysis-and-classification-of-mushrooms

Thank you for waching !!