VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY

UNIVERSITY OF SCIENCE

DECISION TREE WITH SCIKIT-LEARN

COURSE: ARTIFICIAL INTELLIGENCE

TEACHER : Nguyen Ngoc Thao

ASSISTANT TEACHING: Ho Thi Thanh Tuyen

NAME	ID
Pham Minh Xuan	20127395

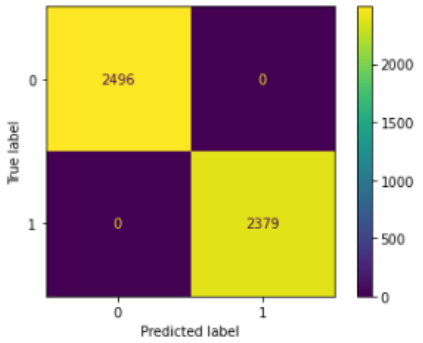
1. Classification report and confusion matrix

- Train 40/ Test 60

Decision Tree Classification

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2496
1	1.00	1.00	1.00	2379
accuracy			1.00	4875
macro avg	1.00	1.00	1.00	4875
weighted avg	1.00	1.00	1.00	4875

Confusion Matrix

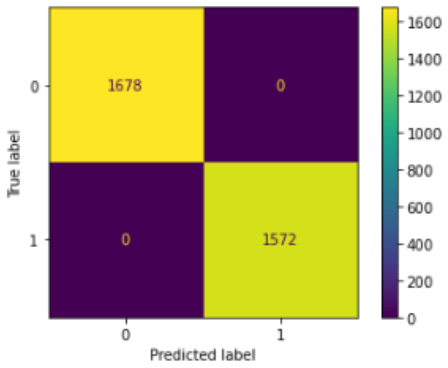


- Train 60/ Test 40

Decision Tree Classification

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1678
1	1.00	1.00	1.00	1572
accuracy			1.00	3250
macro avg	1.00	1.00	1.00	3250
weighted avg	1.00	1.00	1.00	3250

Confusion Matrix

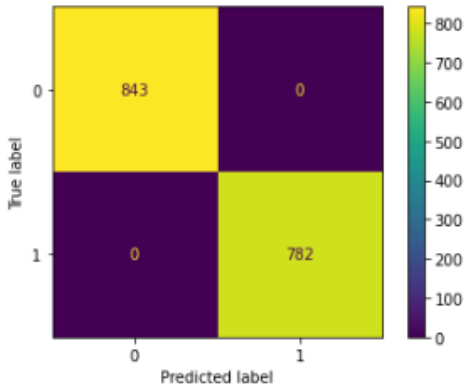


- Train 80/ Test 20

Decision Tree Classification

	precision	recall	f1-score	support
0	1.00	1.00	1.00	843
1	1.00	1.00	1.00	782
accuracy			1.00	1625
macro avg	1.00	1.00	1.00	1625
weighted avg	1.00	1.00	1.00	1625

Confusion Matrix

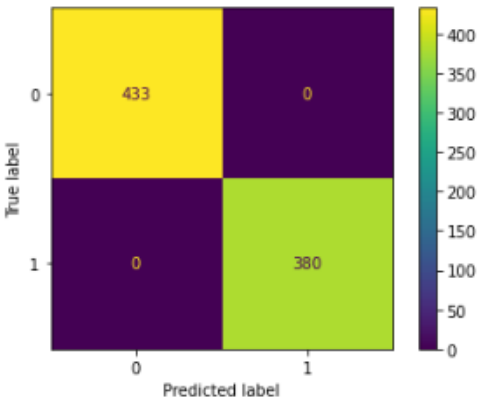


- Train 90/ Test 10

Decision Tree Classification

	precision	recall	f1-score	support
0	1.00	1.00	1.00	433
1	1.00	1.00	1.00	380
accuracy			1.00	813
macro avg	1.00	1.00	1.00	813
weighted avg	1.00	1.00	1.00	813

Confusion Matrix



Classification:

- Precision answers the question in the cases that are predicted to be positive, how many cases are correct? And of course the higher the precision, the better our model.

- Recall measures the rate of correctly predicting positive cases across all samples belonging to the positive group.

  Precision answers the question: of the data points classified by the model into the Positive class, how many data points actually belong to the Positive class. On the other hand, Recall helps us to know how many data points actually in the Positive class are correctly classified by the model in every real data point in the Positive class. Precision and Recall have values in [0,1], the closer these two values are to 1, the more accurate the model is. The higher the precision, the more accurately the grades are classified. A higher recall indicates less omission of correct data points.

- F1-Score is the harmonic mean between precision and recall. It is therefore more representative in the assessment of precision on both precision and recall.
- Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

Confusion matrix

- Confusion matrix provides more information about true classification rates between classes, or helps to detect classes with high misclassification rates thanks to True (False) Positive (Negative) concepts. There are a small number of basic terms which represent for 4 cells in
  the confusion matrix
  - True Positive (TP): the object is in the Positive class, the model assigns the object to the Positive class (correct prediction)
  - True Negative (TN): the object is in the Negative class, the model classifies the object in the Negative class (correct prediction)
  - False Positive (FP): object in the Negative class, the model classifies the object in the Positive class (false prediction) – Type I Error
  - False Negative (FN): the object is in the Positive class, the model classifies the object in the Negative class (false prediction) – Type II Error

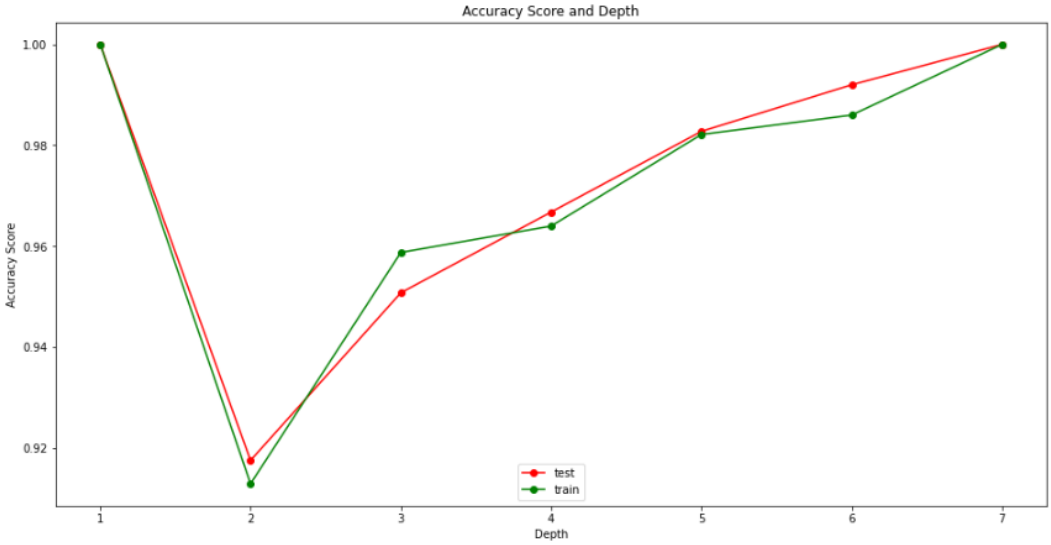
*Comment:

- In 4 different classification , get precision, recall, and f1-score of the same value all of 1.00. Because it has no depth limit.

- Through 4 different models, their performances are slightly similar. The data division is very uniform, there is no situationwhere there are too many elements of a class in the test sets.

2. The depth and accuracy of a decision tree

Accuracy: When building a classification model we will want to know in general what the proportion of correctly predicted cases is out of the total number of cases. That ratio is called accuracy. Accuracy helps us to evaluate the predictive performance of a model on a set of data. The higher the accuracy, the more accurate our model.



max_depth	None	2	3	4	5	6	7
Accuracy	1.0	0.917	0.951	0.967	0.982	0.992	1.0

*Comment:

- According to the above data table, we can see that the accuracy of max_depth= None and max_depth= 7 is the same.
- According to the graph from the point depth = 2, the test line and the train line increase continuously.
- If max_depth is deeper, the model will have higher accuracy and accuracy.
- If the accuracy = 1 then the model is the best, otherwise if the accuracy = 0, the model is the worst.
- The model will be good when the depth is deeper.

3. REFERENCE

- ML model: https://phamdinhkhanh.github.io/2020/08/13/ModelMetric.html
- Scikit-Learn: https://www.tutorialspoint.com/scikit_learn/scikit_learn_modelling_process.htm
- Decision Tree –Depth: https://stackoverflow.com/questions/49289187/decision-tree-sklearn-depth-of-tree-and-accuracy/49289462
- Slide: https://drive.google.com/drive/folders/1OmW_a959zfyCfWP__agipFzfEZX_eV1S
- Decision Tree: https://trituenhantao.io/kien-thuc/decision-tree/
  https://en.wikipedia.org/wiki/Decision_tree
  https://www.youtube.com/watch?v=88rhJ3ow3Us&ab_channel=M%C3%ACAI