UNIVERSITY OF SCIENCE - VNUHCM

FACULTY OF INFORMATION TECHNOLOGY

LAB 02 REPORT

Topic: Decision Tree

Course: Artificial Intelligence

Student:

Võ Hữu Tuấn - 22127439

Lecturer:

Lê Ngọc Thành

Nguyễn Ngọc Thảo

Nguyễn Hải Đăng

Nguyễn Trần Duy Minh

Mục lục

1 Information

1.1 Student information

Name	ID
Võ Hữu Tuấn	22127439

Bảng 1: student information

1.2 Check list

No.	Task	Completion (%)
1	Preparing the data sets	100%
2	Building the decision tree classifiers	100%
3	Evaluating the decision tree classifiers	100%
	Classification report and confusion matrix	
	Comments	
4	The depth and accuracy of a decision tree	100%
	Trees, tables, and charts	
	Comments	
5	Report	100%

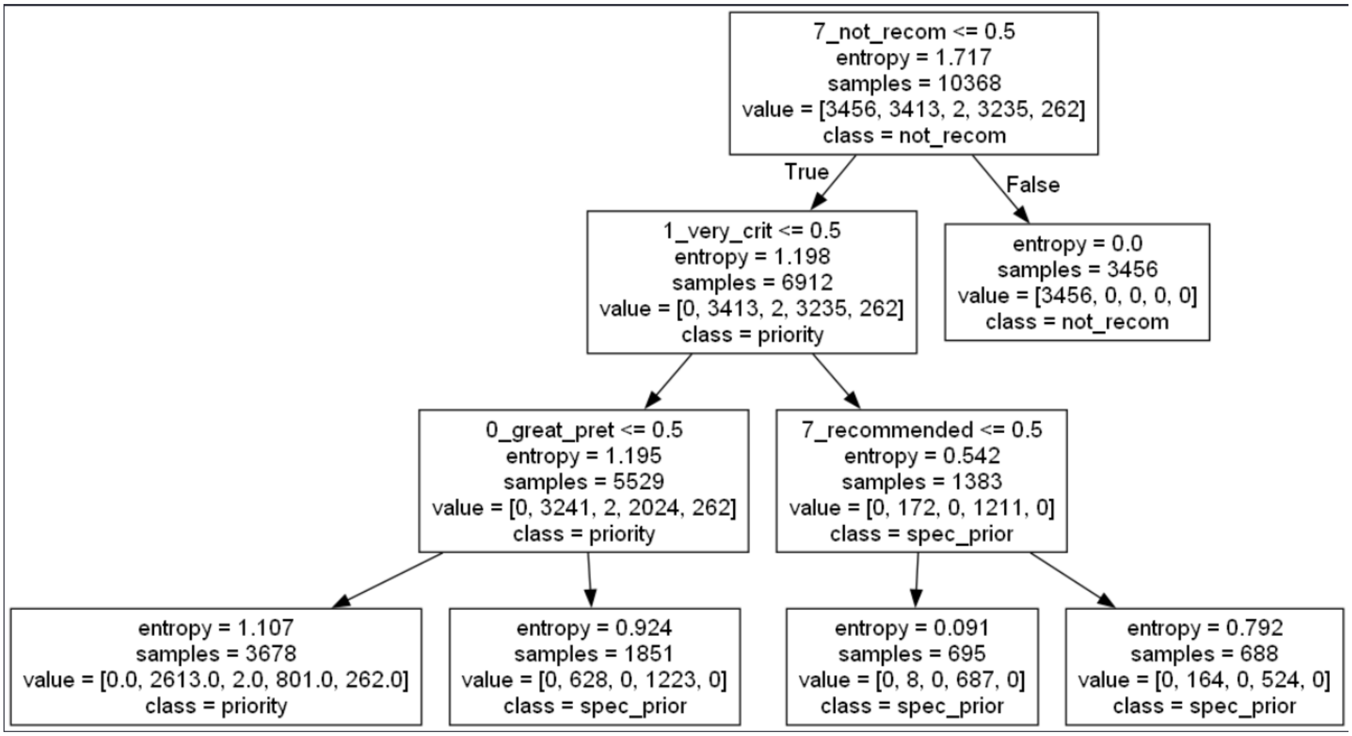
Bảng 2: Check list

2 Requirements

2.1 Preparing the data sets

- Prepare "nursery.data.csv" and read data from it.
- Use `shuffle_and_split_data` function to create training and testing datasets.

2.2 Building the decision tree classifiers



Hình 1: Example of decision tree with ratio 80/20 and max_depth = 2

2.3 Evaluating the decision tree classifiers

2.3.1 Classification report

*** Classification report provides the following information:**

- Precision: The ratio of true positive predictions to the total number of positive predictions (***true positive + false positive***).
- Recall: The ratio of true positive predictions to the total number of actual positive instances (true positive + false negative).
- F1-score: The harmonic mean of precision and recall.
- Support: The number of actual samples in each class.
- Accuracy: The ratio of correct predictions to the total number of samples.

*** Example:**

```
Classification report with 40/60:
              precision    recall  f1-score   support

   not_recom       1.00      1.00      1.00      2592
    priority       0.97      0.98      0.97      2560
   recommend       0.33      1.00      0.50         1
  spec_prior       0.98      0.97      0.97      2426
  very_recom       0.97      0.96      0.96       197

    accuracy                           0.98      7776
   macro avg       0.85      0.98      0.88      7776
weighted avg       0.98      0.98      0.98      7776
```

Hình 2: Classification report with ratio 40/60

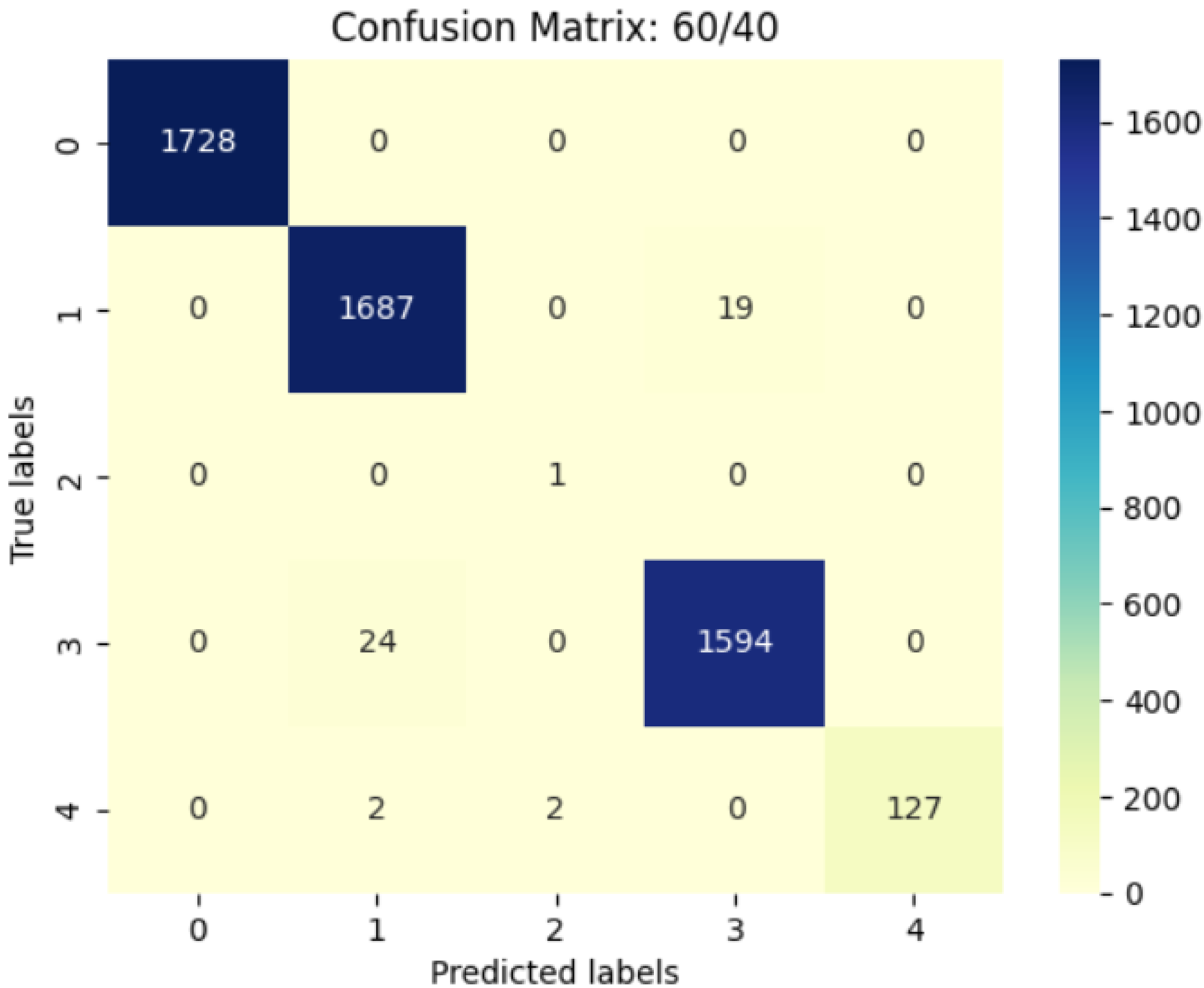
- Comment:
  - The "not_recom" class has perfect precision, recall, and F1-score (=1.00 mean 100%), indicating that the model performs exceptionally well in identifying instances of this class
  - The "priority" and "spec_prior" classes also have high precision, recall, and F1-scores, indicating good performance.
  - The "recommend" class has lower precision but perfect recall, suggesting that the model identifies all instances of this class correctly but may misclassify other instances as "recommend".
  - The "very_recom" class has a slightly lower precision, recall, and F1-score compared to other classes but is still reasonably high.
  - Accuracy: The overall accuracy of the model is reported as 98%, indicating that the model correctly predicts the class label for 98% of the instances in the test set.

2.3.2 Confusion Matrix

*** The confusion matrix provides a visual representation of the performance of the classifiers**.

- Each row of the matrix represents the instances in the actual class:
  - "**not_recom**"
  - **"priority"**
  - **"recommend"**
  - **"spec_prior"**
  - **"very_recom"**
- Each column represents the instances in the predicted class.
- The diagonal elements represent the instances that were correctly classified.
- The off-diagonal elements represent misclassification.

*** Example:**



Hình 3: Confusion Matrix with ratio 60/40

2.4 Evaluating the decision tree classifiers

The decision tree classifier demonstrates strong performance across different train/test ratios

2.5 The depth and accuracy of a decision tree

$\text{Max}_{d}epth$	None	2	3	4	5	6	7
Accuracy	1.00	0.77	0.82	0.84	0.87	0.88	0.91

Bảng 3: Statistical tables

*** Comment:**

- The accuracy score increases steadily as the max_depth parameter increases.
- With max_depth = None, the accuracy reaches 100% (=1.00), indicating potential overfitting due to perfect fitting of the training data.
- With max_depth from 2 to 7, the accuracy steadily improves, which shows that setting an appropriate depth helps prevent overfitting and enhances the model's generalization capability. * In summary, these statistics highlight the importance of controlling the max_depth parameter to strike a balance between model complexity and performance. While a deeper tree may capture intricate patterns in the training data, it risks overfitting, whereas a shallower tree might generalize better to unseen data but could overlook certain nuances in the training data.

Reference

1. Graph Plotting in Python | Set 1

2. Graph Plotting in Python | Set 2

3. Graph Plotting in Python | Set 3

4. Introduction to Scikit-Learn (sklearn) in Python

5. Graphviz