ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

Cơ sở trí tuệ nhân tạo

-Lab Bonus-

Giáo viên : Nguyễn Ngọc Đức

Nguyễn Thị Thu Hằng

Nguyễn Trần Duy Minh

Thành viên nhóm :

STT	MSSV	Họ tên
1	22120059	Trần Minh Đạt
2	22120063	Ngô Phương Đông
3	22120121	Lê Viết Hưng

Hồ Chí Minh ngày 20 tháng 12 năm 2024

Mục lục

1. Phân chia công việc và đánh giá

MSSV	Họ tên	Nội dung	Hoàn thành
22120059	Trần Minh Đạt	The UCI Wine Quality dataset	100%
22120063	Ngô Phương Đông	UCI Breast Cancer Wisconsin (Diagnostic) dataset	100%
22120121	Lê Viết Hưng	UCI Breast Cancer Wisconsin (Original) (Additional Dataset)	100%

2. Chuẩn bị dữ liệu

2.1 The UCI Breast Cancer Wisconsin (Diagnostic) dataset

Số lượng mẫu: 569

Đặc trưng: 30 đặc trưng số

Lớp:

- Lành tính (label = 0)
- Ác tính (label = 1)

Làm sạch dữ liệu

a. Cột ID được loại bỏ vì không liên quan đến phân loại.
b. Các nhãn Diagnosis được ánh xạ thành giá trị nhị phân (M = 1, B = 0).

Chia các tập dữ liệu: Tập dữ liệu được chia thành tập huấn luyện và kiểm tra với các tỷ lệ 40/60, 60/40, 80/20, và 90/10.

➔ Việc chia được thực hiện theo kiểu phân tầng để đảm bảo phân bố lớp được giữ nguyên.

Phân bố lớp: Phân bố lớp được trực quan hóa cho tập dữ liệu gốc và từng tập dữ liệu sau khi chia. Các biểu đồ xác nhận rằng phân bố lớp được duy trì cân bằng

2.2 The UCI Wine Quality dataset

Số lượng mẫu : 4,898 mẫu

Số đặc trưng : 11

Lớp : Từ các giá trị của quality(0-10) chuyển thành 3 lớp High, Standard, Low

Chia các tập dữ liệu: Tập dữ liệu được chia thành tập huấn luyện và kiểm tra với các tỷ lệ 40/60, 60/40, 80/20, và 90/10.

2.3 The UCI Breast Cancer Wisconsin (Original) dataset

Số lượng mẫu : 699

Số đặc trưng : 9

Lớp : có 2 lớp (2 = benign, 4 = malignant)

Chia các tập dữ liệu theo các tỉ lệ 40/60, 60/40, 80/20, and 90/10 (train/test) và trực quan hóa dưới dạng biểu đồ để kiểm tra.

Data : https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

3. Đánh giá mô hình

3.1 The UCI Breast Cancer Wisconsin (Diagnostic) dataset

Theo 40/60:

```
Evaluating Decision Tree for Train/Test Split: 40/60
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.96      0.93       215
           1       0.92      0.83      0.88       127

    accuracy                           0.91       342
   macro avg       0.91      0.90      0.90       342
weighted avg       0.91      0.91      0.91       342

Confusion Matrix:
 [[206   9]
 [ 21 106]]
Accuracy: 0.91
```

→ Mô hình đạt độ chính xác 91% trên 342 mẫu kiểm tra. Đây là mức hiệu suất khá tốt, nhưng có sự chênh lệch khá đáng kể giữa hai lớp

→ Lớp Benign có recall cao (96%) , precision (91%), F1-score (93%) là trung bình điều hòa giữa Precision và Recall, cho thấy mô hình hoạt động rất tốt với lớp này.

→ Lớp Malignant: có precision: 92% đúng trong số các dự đoán ác tính (Malignant), recall: 83% được phân loại đúng trong số các khối u thực sự ác tính, F1-Score: 88% nên hiệu suất thấp hơn so với lớp 0, do Recall của lớp 1 thấp hơn

→ Độ chính xác tổng thể (Accuracy): 91% → 91% dự đoán đúng trên toàn bộ tập kiểm tra.

Kết luận: Mô hình gặp khó khăn hơn trong việc phân loại chính xác các khối u ác tính (Malignant), với 21 trường hợp bị nhầm thành lành tính (Benign). Điều này có thể ảnh hưởng nghiêm trọng trong thực tế vì bỏ sót khối u ác tính có thể dẫn đến hậu quả y khoa nghiêm trọng.

Theo tỉ lệ 60/40:

```
Evaluating Decision Tree for Train/Test Split: 60/40
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.96      0.95       143
           1       0.93      0.91      0.92        85

    accuracy                           0.94       228
   macro avg       0.94      0.93      0.93       228
weighted avg       0.94      0.94      0.94       228

Confusion Matrix:
 [[137   6]
 [  8  77]]
Accuracy: 0.94
```

→ Mô hình đạt độ chính xác 94% trên 228 mẫu kiểm tra → hiệu suất được cải thiện khi tỷ lệ kiểm tra giảm

→ Lớp Benign có recall cao (96%) , precision (94%), F1-score (95%) hiệu suất cao, mô hình xử lý tốt lớp lành tính.

→ Lớp Malignant: có precision: 93% đúng trong số các dự đoán ác tính (Malignant), recall: 91% được phân loại đúng trong số các khối u thực sự ác tính, F1-Score: 92% mô hình làm tốt hơn so với kết quả trước đó, Recall của lớp ác tính đã được cải thiện.

→ Độ chính xác tổng thể (Accuracy): 94% → 94% dự đoán đúng trên toàn bộ tập kiểm tra.

Kết luận: Số lượng nhầm lẫn ở cả hai lớp (6 cho lớp Benign và 8 cho lớp Malignant) đã giảm đáng kể so với kết quả trước. Điều này cho thấy việc chia tập dữ liệu theo tỷ lệ 60/40 có thể đã giúp mô hình học tốt hơn.

So sánh với kết quả trước

- Accuracy: Tăng từ 91% lên 94%.
- F1-Score cho lớp Malignant: Tăng từ 0.88 lên 0.92, phản ánh hiệu suất dự đoán lớp ác tính đã cải thiện.

- Recall của lớp Malignant: Tăng từ 0.83 lên 0.91, cho thấy mô hình bỏ sót ít hơn các trường hợp ác tính.
- Số dự đoán sai (False Negatives) cho lớp 1 giảm từ 21 xuống còn 8.

Kết luận:

- Mô hình hoạt động tốt hơn với tỷ lệ chia tập dữ liệu 60/40, đặc biệt trong việc nhận diện khối u ác tính.
- Độ chính xác tổng thể và các chỉ số đánh giá khác đều được cải thiện so với kết quả trước.
- Tuy nhiên, vẫn có thể tối ưu thêm Recall cho lớp ác tính vì việc bỏ sót (8 trường hợp) vẫn có thể gây nguy hiểm trong bối cảnh y khoa.

Theo 80/20:

```
Evaluating Decision Tree for Train/Test Split: 80/20
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97        72
           1       0.97      0.90      0.94        42

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114

Confusion Matrix:
 [[71  1]
 [ 4 38]]
Accuracy: 0.96
```

→ Độ chính xác đạt 96% trên 114 mẫu kiểm tra, cho thấy mô hình hoạt động tốt trong điều kiện dữ liệu kiểm tra nhỏ hơn

→ Lớp Benign có recall cao (99%) gần như là 100% được phân loại chính xác, precision (95%) trong số các dự đoán lành tính, 95% là đúng, F1-score (97%) hiệu suất cực kỳ tốt đối với lớp lành tính

→ Lớp Malignant: có precision: 97% đúng trong số các dự đoán ác tính (Malignant), recall: 90% được phân loại đúng trong số các khối u thực sự ác tính, F1-Score: 94% hiệu suất dự đoán lớp ác tính khá tốt, tuy nhiên Recall vẫn còn có thể cải thiện.

→ Độ chính xác tổng thể (Accuracy): 96% → 96% dự đoán tỉ lệ cao nhất so với các tỉ lệ trước đó

So sánh với các kết quả trước:

- Accuracy: Cao hơn so với cả hai kết quả trước đó (91% và 94%), đạt 96%.
- F1-Score cho lớp Malignant: Tăng lên 0.94 (cao hơn so với 0.88 ở tỷ lệ 40/60 và 0.92 ở tỷ lệ 60/40).
- Recall của lớp Malignant: Giảm nhẹ so với tỷ lệ 60/40 (từ 0.91 xuống 0.90).
- False Negatives cho lớp Malignant: Ít hơn tỷ lệ 40/60 (từ 21 giảm còn 4) và tỷ lệ 60/40 (giảm từ 8 xuống 4).

Kết luận:

- Hiệu suất cao: Mô hình Decision Tree hoạt động tốt nhất với tỷ lệ chia 80/20, đặc biệt trong việc dự đoán chính xác lớp lành tính (Recall 0.99).
- Tuy nhiên, lớp ác tính: Dù Precision và F1-Score đều cao, Recall 0.90 và 4 trường hợp bị bỏ sót (False Negatives) vẫn đáng lưu ý trong bối cảnh y học.
- Mô hình rất tốt trong việc dự đoán lớp lành tính với chỉ 1 trường hợp bị nhầm lẫn. Tuy nhiên, vẫn có 4 trường hợp ác tính bị bỏ sót, điều này cần lưu ý đặc biệt trong bối cảnh y khoa.

Theo 90/10:

```
Evaluating Decision Tree for Train/Test Split: 90/10
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.97      0.96        36
           1       0.95      0.90      0.93        21

    accuracy                           0.95        57
   macro avg       0.95      0.94      0.94        57
weighted avg       0.95      0.95      0.95        57

Confusion Matrix:
 [[35  1]
 [ 2 19]]
Accuracy: 0.95
```

→ Mô hình đạt độ chính xác 95% trên 57 mẫu kiểm tra. Đây là mức hiệu suất khá tốt khi test với dữ liệu cực kì nhỏ

→ Lớp Benign: mô hình có Recall cao (97%) và Precision ở mức 95%, F1-Score: 0.96 hiệu suất rất cao đối với lớp này.

→ Lớp Malignant: Recall giảm xuống 90%, điều này có thể là dấu hiệu mô hình chưa được kiểm tra đầy đủ trên dữ liệu đa dạng, precision: 95%, F1-score: 96%

→ Độ chính xác tổng thể (Accuracy): 95% → Kết quả rất tốt, chỉ thấp hơn một chút so với tỷ lệ 80/20.

→ Mô hình dự đoán rất chính xác cho cả hai lớp với số lỗi nhầm lẫn rất ít. Tuy nhiên, vẫn còn 2 trường hợp ác tính bị bỏ sót (False Negatives), điều này quan trọng trong ngữ cảnh y khoa.

So sánh với các tỷ lệ trước:

- Accuracy: 95% → Thấp hơn một chút so với tỷ lệ 80/20 (96%) nhưng vẫn cao hơn tỷ lệ 40/60 và 60/40.
- Recall của lớp Malignant (1): 90% → Không thay đổi so với tỷ lệ 80/20 nhưng thấp hơn tỷ lệ 60/40 (91%).
- False Negatives của lớp Malignant: Ít nhất trong tất cả các tỷ lệ trước đó.

Kết luận:

- Hiệu suất cao: Mô hình hoạt động tốt với tỷ lệ chia 90/10, cho kết quả gần tương đương với tỷ lệ 80/20.
- Lớp Malignant: Precision và F1-Score đạt mức cao, nhưng Recall vẫn có thể được cải thiện để giảm thiểu các trường hợp ác tính bị bỏ sót.
- Dữ liệu kiểm tra ít hơn: Với tỷ lệ 90/10, số lượng mẫu kiểm tra nhỏ (57 mẫu) có thể khiến đánh giá hiệu suất kém ổn định hơn so với các tỷ lệ khác.

→ Tỉ lệ train/test của 80/20 đạt độ chính xác cao nhất. Tiếp đến là 90/10, 60/40. 40/60. Từ đó thấy được tỉ lệ dự đoán phần lớn khá cao, có một số ít trường hợp bị nhầm lẫn giữa các lớp

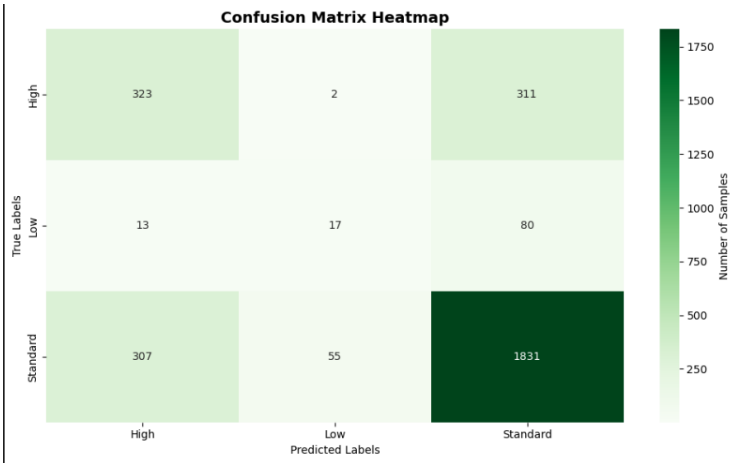
→ Khi phân tích theo độ sâu: Khi tăng độ sâu độ chính xác đạt mức tối đa 95.61% và nó cũng không cải thiện hiệu suất, điều này cho thấy mô hình bắt đầu bị quá khớp.

Kết luận chung:

- Cây quyết định hoạt động tốt nhất với dữ liệu được chia tách cân bằng và sử dụng độ sâu tối ưu.
- Tỷ lệ huấn luyện lớn hơn (80/20, 90/10) cải thiện độ chính xác, nhưng cần đánh đổi với khả năng kiểm tra mô hình.
- Mô hình hoạt động ổn định hơn với lớp Benign. Nhưng hiệu suất của lớp Malignant vẫn thấp hơn so với lớp Benign trên tất cả các tỷ lệ. Đặc biệt, Recall của lớp Malignant dao động từ 83% (40/60) đến 90% (80/20 và 90/10). Điều này có nghĩa là một số mẫu Malignant bị bỏ sót. Trong các vấn đề y tế, việc bỏ sót các trường hợp ung thư là nghiêm trọng, vì vậy cần đặc biệt cải thiện Recall cho lớp Malignant

3.2 The UCI Wine Quality dataset

- **40-60**



Classification Report

	precision	recall	f1-score	support
High	0.50	0.51	0.51	636
Low	0.23	0.15	0.18	110
Standard	0.82	0.83	0.83	2193
accuracy			0.74	2939
macro avg	0.52	0.50	0.51	2939
weighted avg	0.73	0.74	0.74	2939

Confusion Matrix:

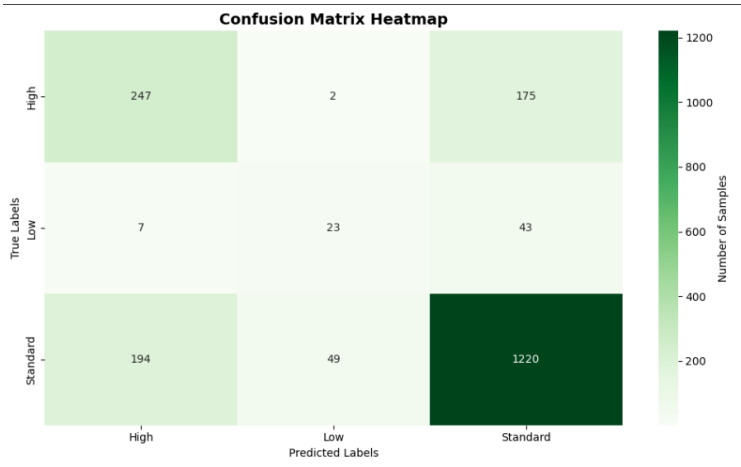
- Dữ liệu tập trung chủ yếu ở lớp Standard với số lượng mẫu rất lớn (1831).
- Lớp Low và High có số lượng mẫu nhỏ hơn và phân tán.
- Có một số nhầm lẫn đáng kể giữa lớp High và Standard (311 và 307) cũng như lớp Low và Standard (80 và 55).

Classification Report:

- Độ chính xác tổng thể: 74%.
- Precision, recall và F1-score của lớp Standard là tốt nhất (0.82-0.83), trong khi lớp Low có hiệu suất rất thấp (precision = 0.23, recall = 0.15).
- Macro average (đánh giá trung bình giữa các lớp) khá thấp (0.51), chỉ ra rằng mô hình không cân bằng tốt giữa các lớp.

=> Nhận xét: Mô hình hoạt động tốt cho lớp chiếm đa số (Standard), nhưng gặp vấn đề nghiêm trọng khi phân loại các lớp Low và High. Điều này có thể do phân bổ dữ liệu không cân bằng.

- **60-40**



Classification Report

	precision	recall	f1-score	support
High	0.55	0.58	0.57	424
Low	0.31	0.32	0.31	73
Standard	0.85	0.83	0.84	1463
accuracy			0.76	1960
macro avg	0.57	0.58	0.57	1960
weighted avg	0.76	0.76	0.76	1960

Confusion Matrix:

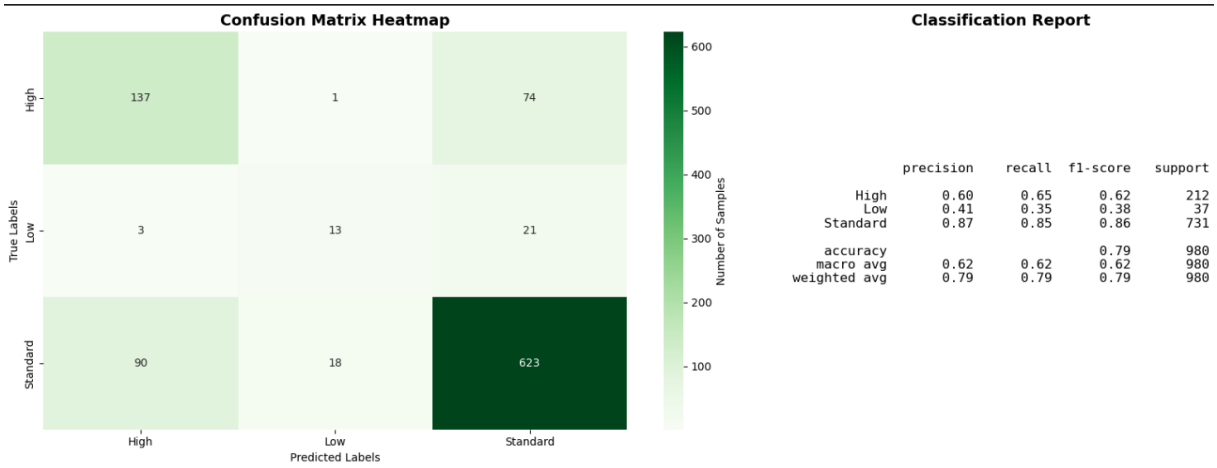
- Số lượng mẫu trong lớp Standard(1220) vẫn lớn hơn nhiều so với các lớp Low(23) và High(247).
- Nhầm lẫn giữa lớp High và Standard (175 và 194) cũng như lớp Standard và Low (43 và 49) khá đáng chú ý.

Classification Report:

- Độ chính xác tổng thể: 76%.
- Precision, recall và F1-score của lớp Standard vẫn tốt nhất (0.85).
- Lớp Low có cải thiện nhẹ so với mô hình đầu tiên nhưng vẫn ở mức thấp (F1-score = 0.31).

=> Nhận xét: Mô hình có sự cải thiện so với tỷ lệ 40-60. Tuy nhiên, lớp Low vẫn cần được cải thiện vì hiệu suất rất thấp. Phân bổ dữ liệu không cân bằng vẫn là một yếu tố chính.

- **80-20**



Confusion Matrix:

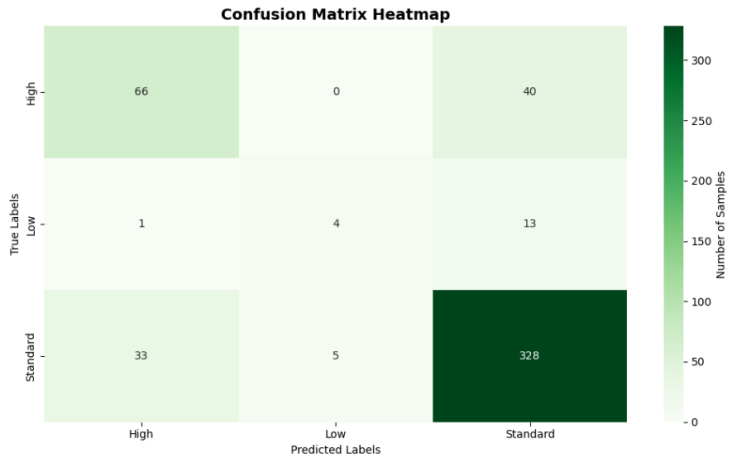
- Phân bổ dữ liệu cân bằng hơn giữa các lớp, với lớp High(212), Low(37), và Standard(731).
- Nhầm lẫn giữa lớp Standard và High (74) ít hơn so với các hình trước.

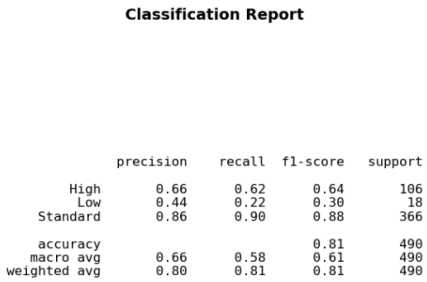
Classification Report:

- Độ chính xác tổng thể: 79%.
- Precision, recall và F1-score của lớp Standard vẫn cao (0.87,0.85) và lớp Low vẫn có hiệu suất thấp nhất (F1-score = 0.38).
- Macro average và weighted average đều tăng so với hai mô hình trước.

=> Nhận xét: Phân bổ dữ liệu cân bằng hơn giúp mô hình hoạt động ổn định hơn. Tuy nhiên, hiệu suất của lớp Low vẫn thấp và có thể cải thiện thêm.

- **90-10**





Confusion Matrix

- Phân bổ dữ liệu cân bằng hơn giữa các lớp, với lớp High(106), Low(18), và Standard(366).
- Nhầm lẫn giữa lớp Standard và High (40) là đáng kể nhưng ít hơn so với các mô hình khác.

Classification Report

- Độ chính xác tổng thể: 81%.
- Precision, recall, và F1-score của lớp Standard vẫn cao, cho thấy mô hình phân loại tốt với lớp này.
- Lớp Low có hiệu suất thấp nhất (F1-score = 0.30), có thể do dữ liệu không đủ (18 mẫu).
- Macro average và weighted average đạt mức tốt (Macro F1-score = 0.61, Weighted F1-score = 0.81).

=> Nhận xét:Phân bổ dữ liệu tương đối ổn, nhưng lớp Low vẫn cần được cải thiện thêm.Hiệu suất tốt của lớp Standard cho thấy mô hình hoạt động hiệu quả hơn.

3.3 The UCI Breast Cancer Wisconsin (Original) dataset

- **Tỉ lệ 40/60 :**

	precision	recall	f1-score	support
Benign	0.94	0.97	0.95	275
Malignant	0.94	0.88	0.91	145
accuracy			0.94	420
macro avg	0.94	0.92	0.93	420
weighted avg	0.94	0.94	0.94	420

➔ Mô hình đạt độ chính xác tổng thể khá cao 94% trên 420 mẫu kiểm tra

➔ Lớp Benign : chỉ số recall của lớp Benign cao (97%) cho thấy mô hình hoạt động tốt trong việc phát hiện các trường hợp Benign.

➔ Lớp Malignant : chỉ số recall còn thấp (88%) cho thấy mô hình còn hoạt động chưa tốt trong việc dự đoán về các trường hợp là Malignant.

Kết luận : Cần phải cải thiện hơn về lớp Malignant vì sẽ rất nguy hiểm nếu bỏ sót nhiều như vậy.

- **Tỉ lệ 60/40 :**

	precision	recall	f1-score	support
Benign	0.93	0.96	0.94	183
Malignant	0.91	0.87	0.89	97
accuracy			0.93	280
macro avg	0.92	0.91	0.92	280
weighted avg	0.92	0.93	0.92	280

➔ Độ chính xác tổng thể 93% trên 280 mẫu kiểm thử (thấp hơn so với trước)

➔ Lớp Benign : tỉ lệ chính xác cho các trường hợp Benign (93%) khá thấp tuy nhiên tỉ lệ bỏ sót các trường hợp là benign khá thấp, điều này có thể gây yên tâm cho người được chuẩn đoán.

➔ Lớp Malignant : tỉ lệ bỏ sót các trường hợp là Malignant cao 13%, điều này rất nguy hiểm cho các bệnh nhân, chỉ số f1-score cũng khá thấp chỉ ở mức 89%

Kết luận : Mô hình bỏ sót nhiều trường hợp là malignant, hoạt động chưa tốt

- **Theo tỉ lệ 80/20**

	precision	recall	f1-score	support
Benign	0.94	0.97	0.95	92
Malignant	0.93	0.88	0.90	48
accuracy			0.94	140
macro avg	0.94	0.92	0.93	140
weighted avg	0.94	0.94	0.94	140

➔ Độ chính xác của mô hình ở mức 94% trên 140 mẫu
➔ Lớp Benign : việc dự đoán và nhận diện các trường hợp benign khá cao
➔ Lớp Malignant : việc dự đoán và nhận diện các trường hợp là malignant còn ở mức khá thấp , vẫn bỏ sót nhiều trường hợp dẫn đến nguy hiểm cho bệnh nhân

Kết luận : Mô hình vẫn chỉ dừng lại ở mức khá tốt

- **Theo tỉ lệ 90/10**

	precision	recall	f1-score	support
Benign	1.00	0.96	0.98	46
Malignant	0.92	1.00	0.96	24
accuracy			0.97	70
macro avg	0.96	0.98	0.97	70
weighted avg	0.97	0.97	0.97	70

➔ Độ chính xác cao ở mức 97% trên 70 mẫu

➔ Lớp Benign : f1-score ở mức 98% cho thấy cả việc dự đoán lẫn nhận diện các trường hợp đều ở mức tốt

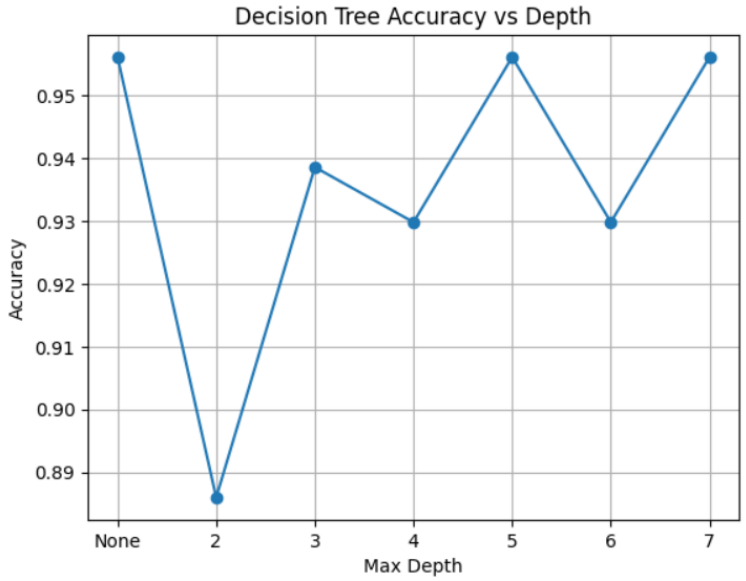
➔ Lớp Malignant : chỉ số precision giảm so với các trường hợp trước tuy nhiên recall ở mức chính xác nhất 100%

Kết luận : Mô hình hoạt động tốt nhưng do dữ liệu kiểm tra nhỏ nên không đủ để đánh giá tính tổng quát của mô hình, nên sử dụng tỉ lệ train/test ở mức thấp hơn

4. Ảnh hưởng của độ sâu (80/20)

4.1 The UCI Breast Cancer Wisconsin (Diagnostic) dataset

Max_depth	None	2	3	4	5	6	7
Accuracies	0.9561	0.8860	0.9386	0.9298	0.9561	0.9298	0.9561



- max-depth=None: Không giới hạn độ sâu (None) cho thấy mô hình đạt được độ chính xác cao nhất (95.61%). Điều này cho thấy cây quyết định tận dụng hết tiềm năng phân chia dữ liệu để đạt độ chính xác tối đa. Tuy nhiên, điều này có thể đi kèm nguy cơ overfitting.
- max-depth=2: Khi giới hạn độ sâu ở mức 2, độ chính xác giảm xuống 88.60%. Điều này chứng tỏ độ sâu này không đủ để cây quyết định tạo ra các nút phân chia hiệu quả, dẫn đến underfitting.
- max-depth=3: Độ chính xác cải thiện đáng kể so với độ sâu 2, đạt 93.86%. Độ sâu này có vẻ cân bằng hơn, cho phép cây phân chia dữ liệu hiệu quả hơn mà không quá phức tạp.

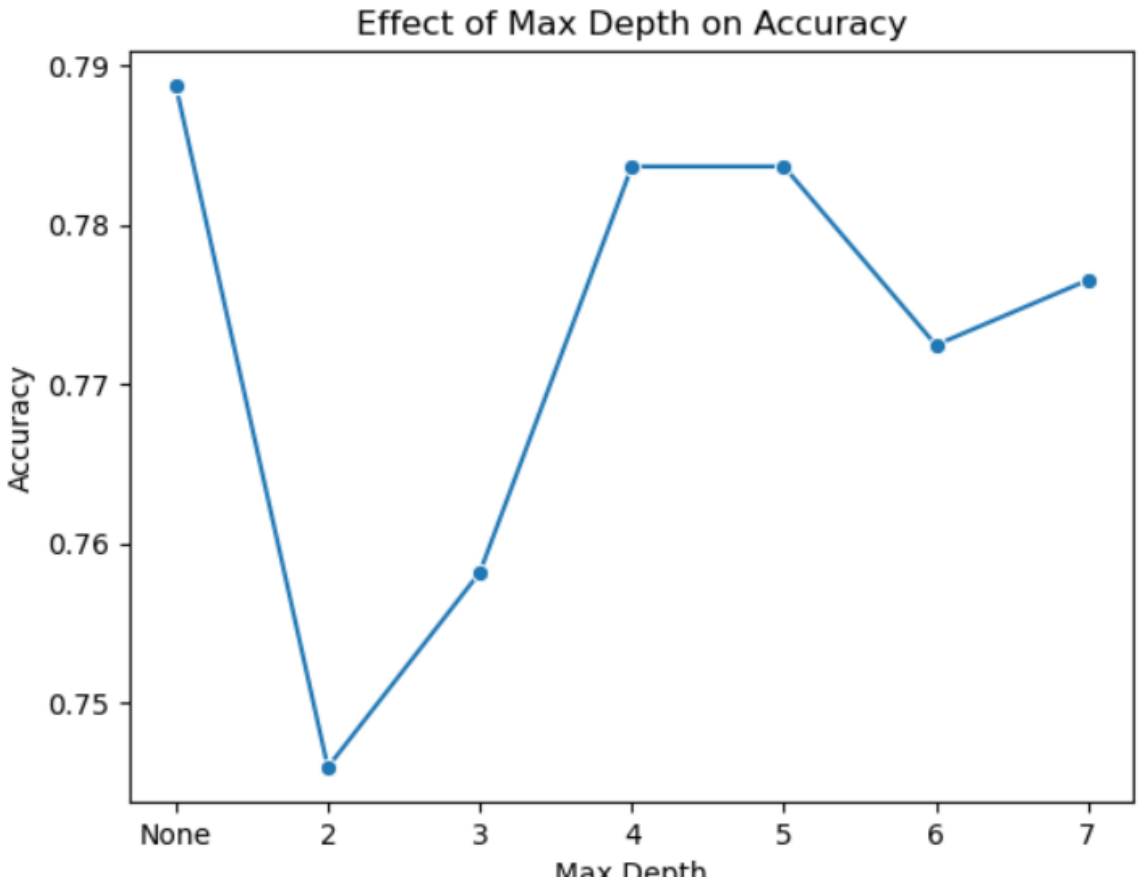
- max-depth=4: Độ chính xác giảm nhẹ còn 92.98%. Điều này có thể cho thấy độ sâu này không mang lại cải thiện đáng kể về độ phân chia dữ liệu so với max_depth = 3.
- max-depth=5: Tương tự max_depth = None, độ chính xác đạt 95.61%. Điều này có thể cho thấy ở độ sâu này, cây quyết định đủ khả năng phân chia toàn diện mà không bị overfitting nhiều.
- max-depth=6: Độ chính xác giảm xuống 92.98%. Có thể cây đã thêm các nhánh phức tạp không thực sự cải thiện phân loại mà gây nhiễu.
- max-depth=7: Độ chính xác quay lại mức 95.61%. Tương tự các độ sâu None và 5, điều này gợi ý rằng độ sâu lớn hơn có thể giúp mô hình bao phủ tốt hơn các mẫu phức tạp trong tập dữ liệu.

Kết luận chung:

- Cây quyết định hoạt động rất tốt khi độ sâu đủ lớn để nắm bắt tất cả mối quan hệ trong dữ liệu.
- max_depth = 3 cũng cho thấy độ chính xác tốt (93.86%) trong khi giúp hạn chế phức tạp hóa mô hình.
- max_depth = 2 không đủ khả năng phân chia dữ liệu hiệu quả, gây ra underfitting.
- max_depth = 6 có dấu hiệu giảm độ chính xác, có thể liên quan đến overfitting nhẹ.
- Các giá trị max_depth = 5 hoặc max_depth = 7 có thể là lựa chọn tốt nhất, đảm bảo độ chính xác cao và tính tổng quát của mô hình. Điều này giúp giảm nguy cơ overfitting mà vẫn đạt hiệu suất cao.

4.2 The UCI Wine Quality dataset

Max_depth	None	2	3	4	5	6	7
Accuracies	0.789	0.746	0.758	0.784	0.784	0.772	0.776



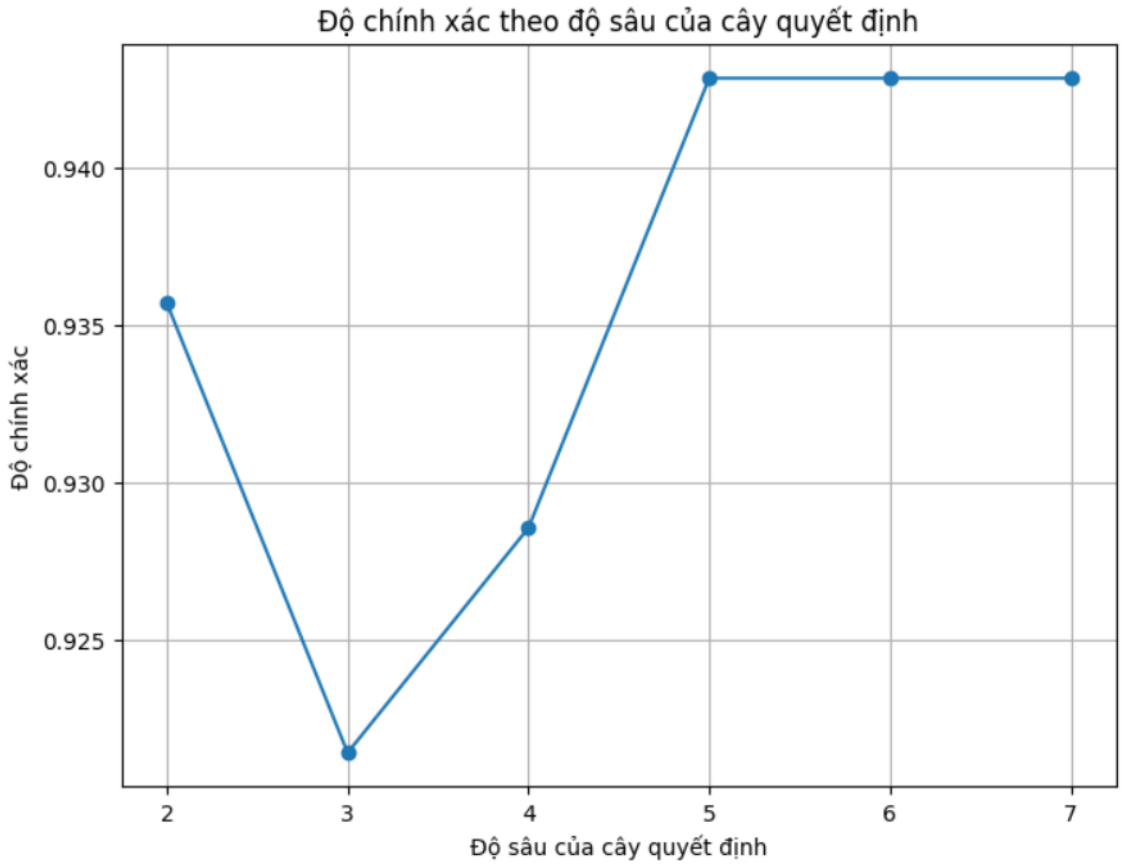
Nhận xét :

- Khi Max Depth là "None" (không giới hạn), độ chính xác đạt giá trị cao nhất là 0.789.
- Độ chính xác thấp nhất xảy ra khi Max Depth bằng 2 (0.746).
- Giá trị Max Depth từ 4 đến 5 cho kết quả độ chính xác tương đối ổn định ở mức 0.784.
- Xu hướng giảm mạnh ban đầu: Khi Max Depth thay đổi từ "None" xuống 2, độ chính xác giảm đáng kể.
- Tăng dần: Sau Max Depth = 2, độ chính xác tăng đều đặn và đạt giá trị cao ở Max Depth = 4, 5.
- Sụt giảm nhẹ và tăng trở lại: Khi Max Depth vượt qua 5, độ chính xác giảm và sau đó có dấu hiệu cải thiện ở Max Depth = 7.

=> Max Depth có tác động rõ ràng đến độ chính xác của mô hình.Việc không giới hạn độ sâu hoặc đặt giới hạn hợp lý (4-5) có vẻ là tối ưu trong trường hợp này.Quá lớn hoặc quá nhỏ giá trị Max Depth đều dẫn đến sự giảm độ chính xác, có thể do hiện tượng underfitting (do Max Depth nhỏ) hoặc overfitting (do Max Depth lớn).

4.3 The UCI Breast Cancer Wisconsin (Original) dataset

Max_depth	None	2	3	4	5	6	7
Accuracies	0.9500	0.9357	0.9214	0.9286	0.9429	0.9429	0.9429



- Max_depth = None : độ chính xác cao nhất đạt 95%
- Khi tăng độ sâu từ 2 đến 7 :
    - Độ chính xác giảm xuống thấp nhất **0.9214** tại độ sâu 3
    - Sau đó tăng dần và đạt giá trị ổn định từ độ sâu 5 trở đi
    - Độ sâu từ 5-7 không đổi

➔ Kết luận : độ sâu = None cho kết quả tốt nhất tuy nhiên có nguy cơ overfitting nên, độ sâu 5 là độ sâu lí tưởng cho cây quyết định, mô hình đơn giản hơn độ sâu 6, 7 nhưng vẫn cho kết quả tương tự.

5. Tài liệu tham khảo

https://gist.github.com/pb111/af439e4affb1dd94879579cfd6793770

https://scikit-learn.org/1.5/auto_examples/release_highlights/plot_release_highlights_1_5_0.html#sphx-glr-auto-examples-release-highlights-plot-release-highlights-1-5-0-py

https://stackoverflow.com/questions/58022382/classification-report-parameters-for-decision-trees-precision-recall-f1-score

https://stackoverflow.com/questions/42621190/display-this-decision-tree-with-graphviz?utm_source=chatgpt.com

https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original