

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN



PROJECT BONUS – DECISION TREE

MÔN: CƠ SỞ TRÍ TUỆ NHÂN TẠO

Giáo viên hướng dẫn: Nguyễn Ngọc Đức

Sinh viên thực hiện: Lê Quang Vĩnh Quyền (*)

Hồ Quang Sang

Võ Tuấn Tài

Nguyễn Minh Tâm

Lớp: CQ2022/22

Mục lục

1.	Mức độ hoàn thành	1
2.	Wine Quality dataset	1
2.1.	Tập dữ liệu	1
2.2.	Chuẩn bị dữ liệu	1
2.3.	Xây dựng mô hình cây quyết định	3
2.4.	Đánh giá mô hình cây quyết định	5
2.5.	Ảnh hưởng của độ sâu đối với độ chính xác của cây quyết định	12
3.	Breast Cancer dataset	16
3.1.	Tập dữ liệu	16
3.2.	Chuẩn bị dữ liệu	16
3.3.	Xây dựng mô hình cây quyết định	18
3.4.	Đánh giá mô hình cây quyết định	22
3.5.	Ảnh hưởng của độ sâu đối với độ chính xác của cây quyết định	28
4.	Additional dataset (Mushroom)	35
4.1.	Tập dữ liệu	35
4.2.	Chuẩn bị dữ liệu	35
4.3.	Xây dựng mô hình cây quyết định	37
4.4.	Đánh giá mô hình cây quyết định	42
4.5.	Ảnh hưởng của độ sâu đối với độ chính xác của cây quyết định	46
5.	So sánh trên 3 tập dữ liệu	52
5.1	Đặc điểm dữ liệu	52
5.2	Độ phức tạp của mô hình	52
5.3	Hiệu năng thực tế	52
5.4	So sánh tổng quan	53

Báo cáo

1. Mức độ hoàn thành

MSSV	Họ và tên	Công việc	Hoàn thành
	Lê Quang Vĩnh Quyền	Additional dataset (Mushroom)	100%
	Hồ Quang Sang	Compare 3 dataset	100%
	Võ Tuấn Tài	Wine Quality dataset	100%
	Nguyễn Minh Tâm	Breast Cancer dataset	100%

2. Wine Quality dataset

2.1. Tập dữ liệu

Nguồn: [Wine Quality - UCI Machine Learning Repository](#)

Thông tin:

- Số mẫu: 4898
- Số đặc trưng: 11 (các tính chất hóa học)
- Mục tiêu: Phân loại rượu thành ba nhóm chất lượng:
 - o Chất lượng thấp: Nhóm 0-4
 - o Chất lượng trung bình: Nhóm 5-6
 - o Chất lượng cao: Nhóm 7-10

2.2. Chuẩn bị dữ liệu

Đọc dữ liệu từ tệp CSV bằng pandas.

Tiền xử lý:

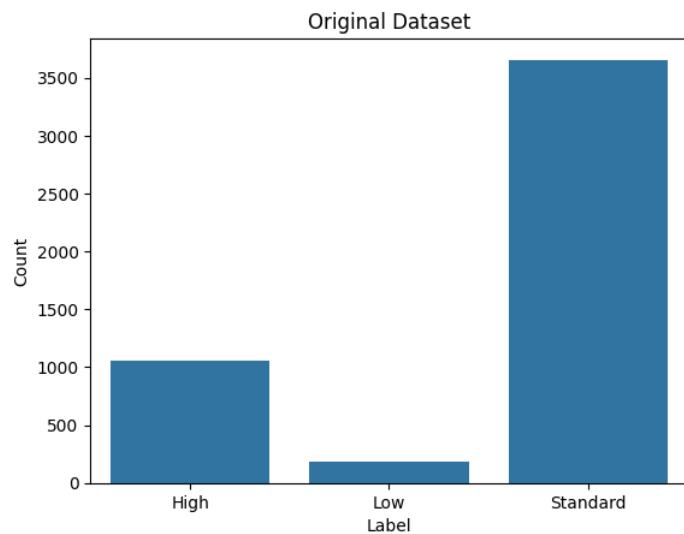
- Nhóm nhãn chất lượng thành ba nhóm: chất lượng thấp (0-4), chất lượng trung bình (5-6), và chất lượng cao (7-10).
- Chuẩn hóa các đặc trưng bằng MinMaxScaler hoặc StandardScaler.

Chia dữ liệu thành các tập:

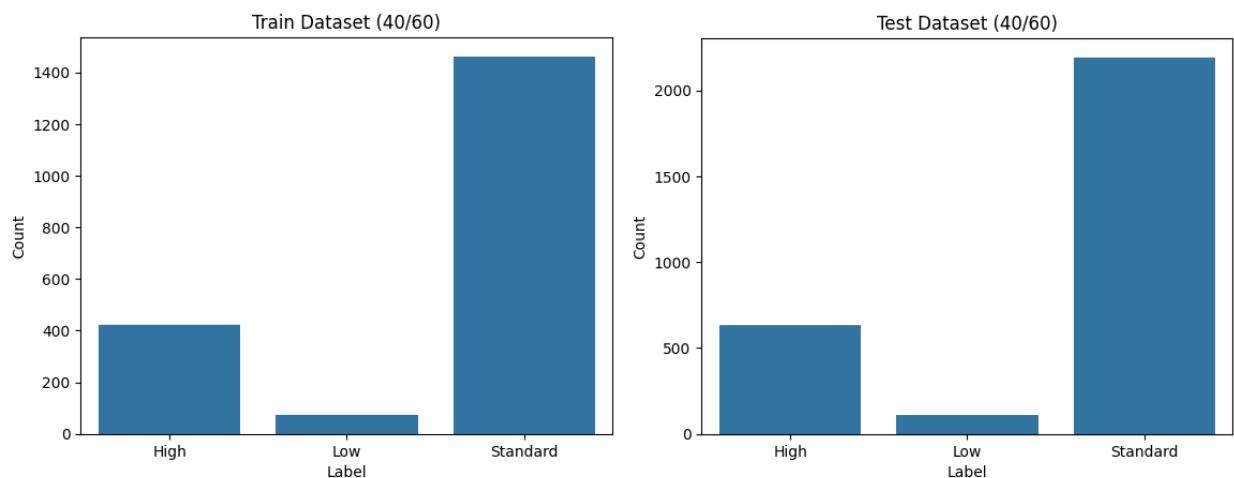
- Chia thành các tập train/test theo tỷ lệ 40/60, 60/40, 80/20, và 90/10.
- Sử dụng StratifiedShuffleSplit để giữ nguyên phân phối nhãn.

Trực quan hóa: Tạo biểu đồ phân phối các nhãn chất lượng (thấp/trung bình/cao).

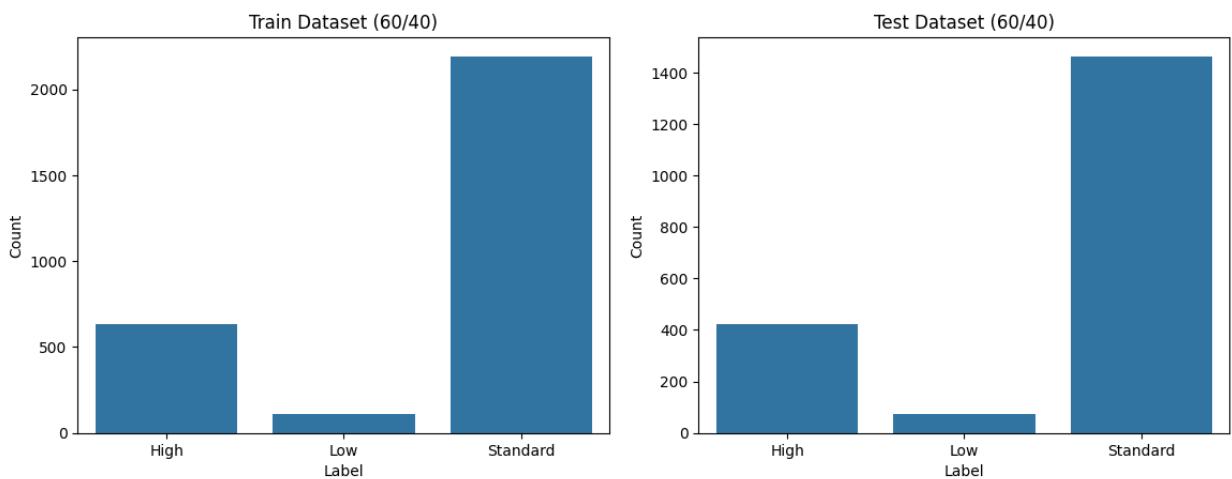
Dữ liệu gốc



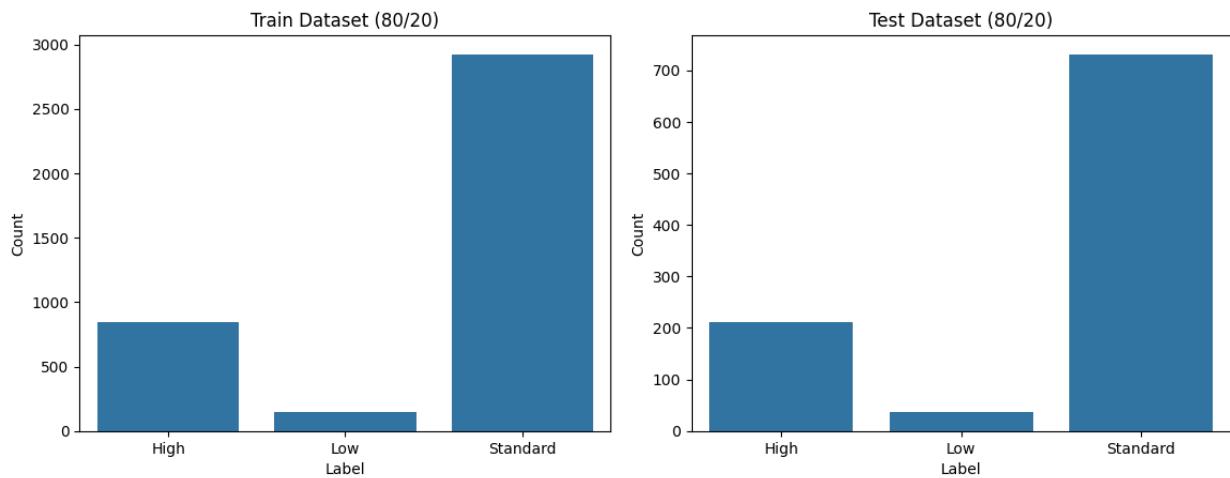
Tập train/test theo các tỷ lệ 40/60



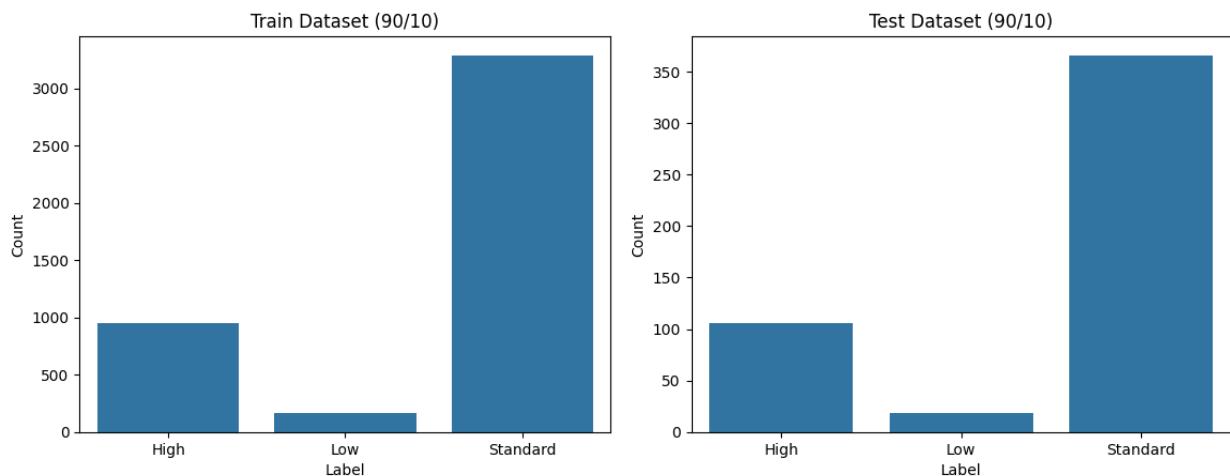
Tập train/test theo các tỷ lệ 60/40



Tập train/test theo các tỷ lệ 80/20



Tập train/test theo các tỷ lệ 90/10



2.3. Xây dựng mô hình cây quyết định

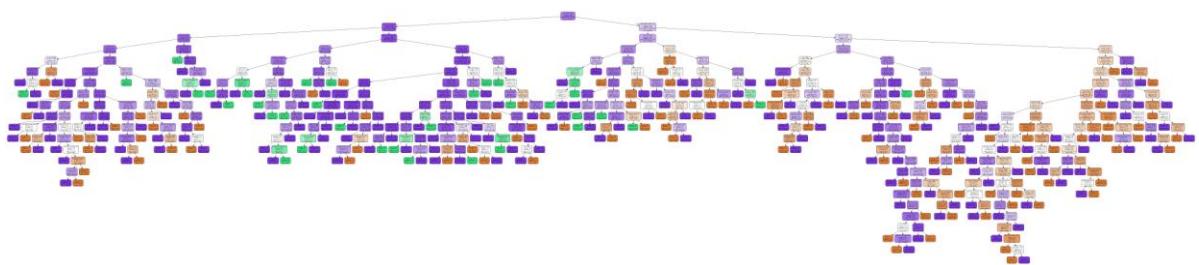
Công cụ sử dụng: DecisionTreeClassifier từ thư viện scikit-learn.

Thông số quan trọng:

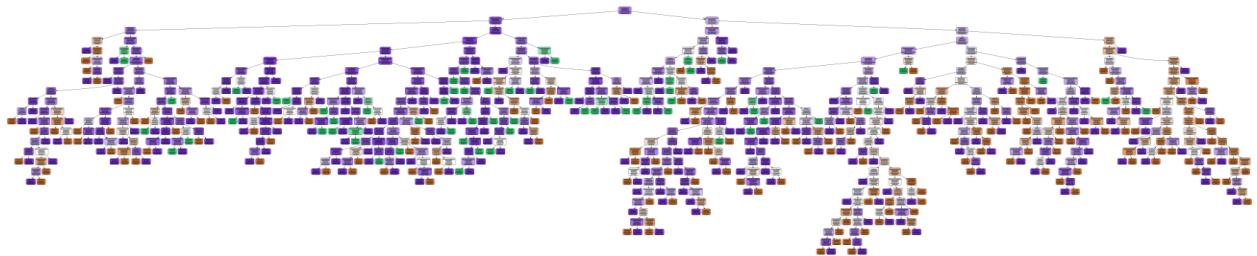
- Tiêu chí phân tách: Information Gain (entropy).
- Hiển thị cây quyết định bằng Graphviz.

Kết quả mong đợi: Cây quyết định được hiển thị cho từng tỷ lệ train/test khác nhau.

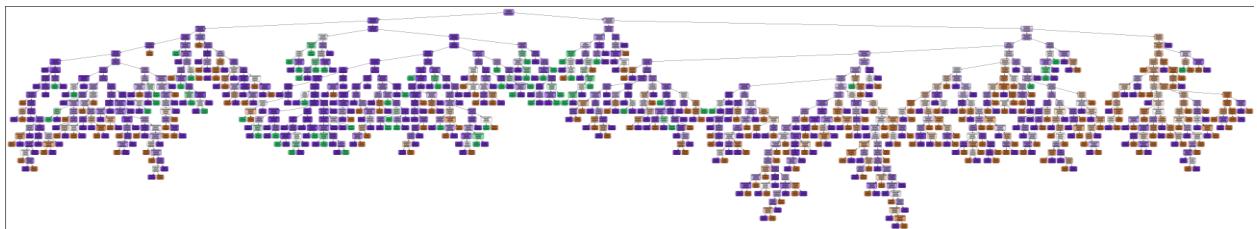
Tỷ lệ tập train/test là 40/60



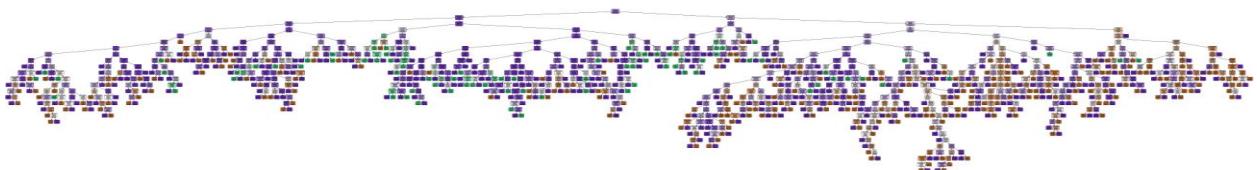
Tỷ lệ tập train/test là 60/40



Tỷ lệ tập train/test là 80/20



Tỷ lệ tập train/test là 90/10



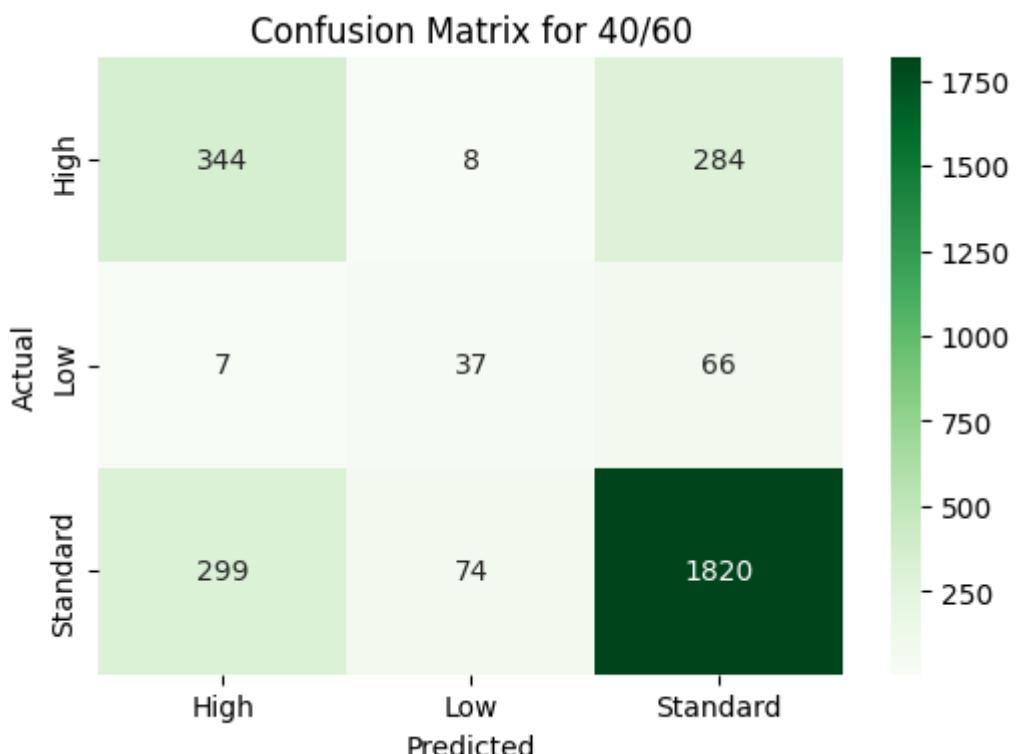
2.4. Đánh giá mô hình cây quyết định

Classification Report cho tỷ lệ 40/60

	precision	recall	f1-score	support
High	0.52	0.53	0.53	636
Low	0.22	0.20	0.21	110
Standard	0.83	0.83	0.83	2193
accuracy			0.79	2939
macro average	0.52	0.52	0.52	2939
weighted average	0.74	0.74	0.74	2939

- Hiệu suất của các lớp:
 - + Lớp High: Độ chính xác (Precision): 0.52, độ nhạy (Recall): 0.53 → Mô hình có khả năng phân loại lớp "High" tương đối tốt, tuy nhiên vẫn tồn tại sự nhầm lẫn đáng kể với các lớp khác.
 - + Lớp Low: Độ chính xác: 0.22, độ nhạy: 0.20 → Hiệu suất của lớp "Low" rất thấp, đây là lớp bị dự đoán sai nhiều nhất.
 - + Lớp Standard: Độ chính xác: 0.83, độ nhạy: 0.83 → Lớp "Standard" được dự đoán chính xác nhất, với cả độ chính xác và độ nhạy đều đạt mức cao.
- Ma trận nhầm lẫn (Confusion Matrix):
 - + Lớp Standard: Dự đoán chính xác 1815 mẫu. Tuy nhiên, có 308 mẫu bị nhầm lẫn thành lớp "High" và 70 mẫu nhầm thành lớp "Low".
 - + Lớp High: Chỉ có 339 mẫu được dự đoán đúng. Tồn tại 289 mẫu bị nhầm lẫn thành lớp "Standard".
 - + Lớp Low: Chỉ có 22 mẫu được dự đoán chính xác. Tới 80 mẫu bị nhầm thành lớp "Standard".
- Tổng quan hiệu suất:
 - + Độ chính xác toàn cục (Accuracy): 0.79. Mô hình đạt độ chính xác khá cao trên toàn bộ tập dữ liệu.

- + Trung bình macro (Macro Average): Precision, Recall, và F1-Score đều là 0.52 → Điều này phản ánh hiệu suất trung bình trên từng lớp, không tính đến độ chênh lệch kích thước giữa các lớp.
- + Trung bình trọng số (Weighted Average): Precision, Recall, và F1-Score đều là 0.74 → Mô hình tập trung tốt hơn vào lớp "Standard", lớp có số lượng lớn nhất, dẫn đến trung bình trọng số cao hơn.
- Nhận xét:
 - + Hiệu suất tốt nhất: Lớp "Standard" được dự đoán với hiệu suất rất cao.
 - + Nhược điểm lớn nhất: Lớp "Low" có độ chính xác và độ nhạy thấp nhất, gây ra tỷ lệ dự đoán sai rất cao.
 - + Kiểm tra lại các đặc trưng đầu vào để đảm bảo thông tin có đủ để phân biệt rõ giữa các lớp, đặc biệt là giữa "Low" và các lớp khác, có thể bằng cách sử dụng các kỹ thuật cân bằng dữ liệu



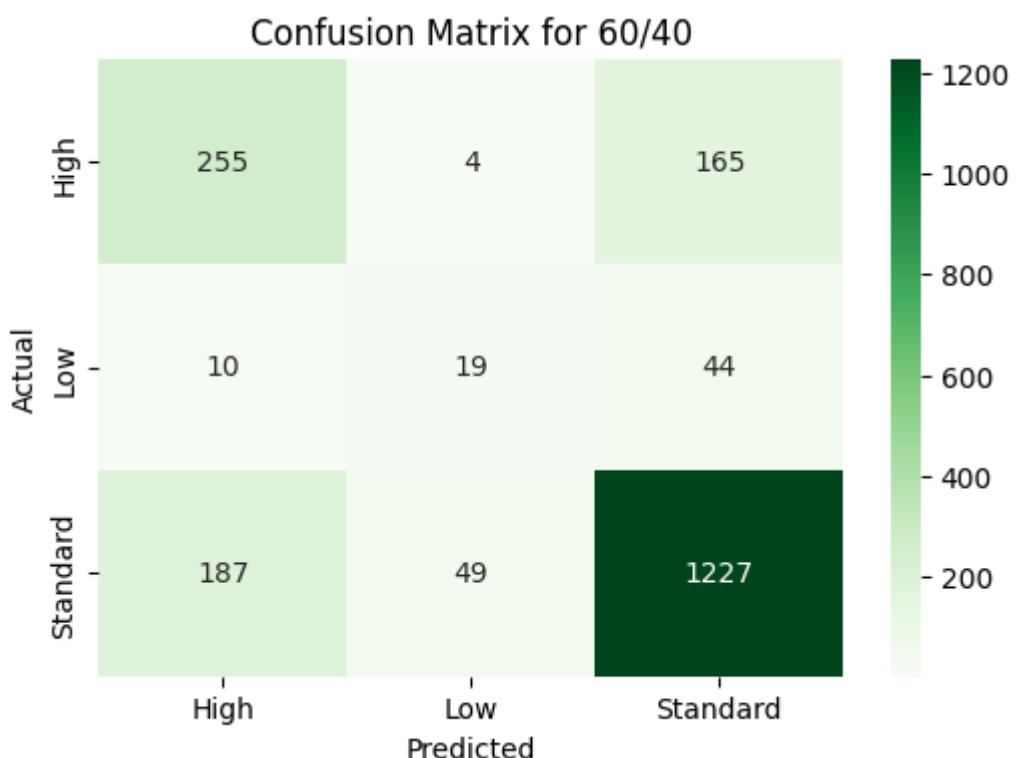
Classification Report cho tỷ lệ 60/40

- Hiệu suất của các lớp:
 - + Lớp High: Độ chính xác (Precision): 0.58, độ nhạy (Recall): 0.61 → Có sự cải thiện nhẹ so với mô hình 40/60, tuy nhiên lớp "High" vẫn gặp khó khăn trong việc phân loại chính xác và vẫn bị nhầm lẫn với các lớp khác.
 - + Lớp Low: Độ chính xác: 0.29, độ nhạy: 0.33 → Hiệu suất của lớp "Low" vẫn thấp và bị dự đoán sai nhiều, không có sự cải thiện rõ rệt so với mô hình trước.
 - + Lớp Standard: Độ chính xác: 0.86, độ nhạy: 0.85 → Lớp "Standard" tiếp tục được dự đoán chính xác nhất, với độ chính xác và độ nhạy cao, thể hiện sự ổn định trong phân loại.
- Ma trận nhầm lẫn (Confusion Matrix):
 - + Lớp Standard: Dự đoán đúng 1,233 mẫu. Tuy nhiên, có 178 mẫu bị nhầm thành lớp "High" và 52 mẫu nhầm thành lớp "Low".
 - + Lớp High: Chỉ có 257 mẫu được dự đoán đúng. Tồn tại 159 mẫu bị nhầm thành lớp "Standard".
 - + Lớp Low: Chỉ có 24 mẫu được dự đoán đúng. 41 mẫu bị nhầm thành lớp "Standard".
- Tổng quan hiệu suất:
 - + Độ chính xác toàn cục (Accuracy): 0.77. Mô hình đạt độ chính xác cao nhưng vẫn có nhiều mẫu bị nhầm lẫn, đặc biệt là với lớp "Low".

	precision	recall	f1-score	support
High	0.58	0.61	0.59	424
Low	0.29	0.33	0.31	73
Standard	0.86	0.85	0.84	1463
accuracy			0.77	1960
macro average	0.58	0.59	0.58	1960
weighted average	0.78	0.77	0.78	1960

- + Trung bình macro (Macro Average): Precision: 0.58, Recall: 0.59, F1-Score: 0.58 → Mức độ trung bình trên các lớp, cho thấy hiệu suất phân loại chung không hoàn hảo, đặc biệt đối với các lớp không phổ biến như "Low".

- + Trung bình trọng số (Weighted Average): Precision: 0.78, Recall: 0.77, F1-Score: 0.78
→ Do lớp "Standard" chiếm số lượng lớn nhất trong dữ liệu, trung bình trọng số có giá trị cao hơn, cho thấy mô hình tập trung tốt vào lớp này.
- Nhận xét:
 - + Hiệu suất tốt nhất: Lớp "Standard" có hiệu suất cao nhất và được phân loại chính xác nhất.
 - + Nhược điểm lớn nhất: Lớp "Low" vẫn có hiệu suất rất kém, với độ chính xác và độ nhạy thấp, và bị nhầm lẫn mạnh mẽ với lớp "Standard".
 - + Mặc dù có một số cải thiện nhẹ so với mô hình trước (40/60), mô hình vẫn gặp khó khăn lớn trong việc phân loại chính xác lớp "Low". Cần phải tập trung cải thiện việc phân loại lớp này, cũng như giảm nhầm lẫn giữa lớp "High" và lớp "Standard". Mô hình có thể đạt hiệu suất tốt hơn nếu áp dụng các kỹ thuật cân bằng dữ liệu và tối ưu hóa các siêu tham số.

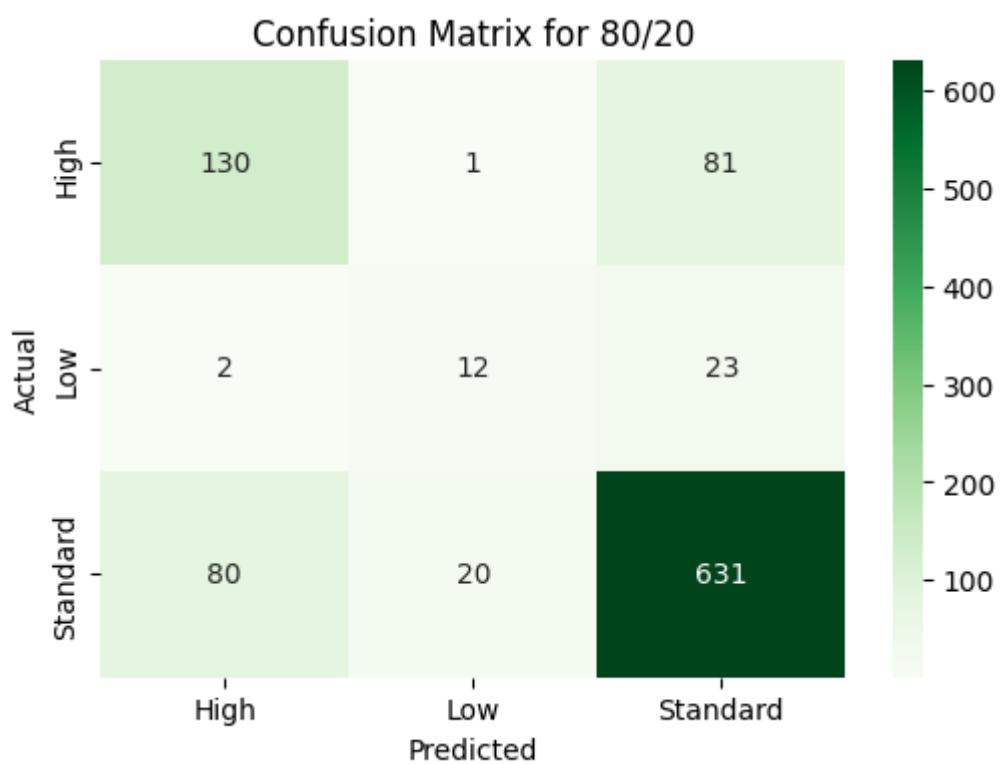


Classification Report cho tỷ lệ 80/20

	precision	recall	f1-score	support
High	0.56	0.58	0.57	212
Low	0.42	0.42	0.41	37
Standard	0.86	0.85	0.85	731
accuracy			0.78	980
macro average	0.61	0.61	0.61	980
weighted average	0.78	0.78	0.78	980

- Hiệu suất của các lớp:
 - + Lớp High: Độ chính xác (Precision): 0.56, độ nhạy (Recall): 0.58 → Có sự cải thiện đáng kể so với các mô hình trước, tuy nhiên, vẫn còn nhầm lẫn giữa lớp "High" và các lớp khác. Cần tiếp tục cải thiện độ chính xác để phân loại lớp "High" rõ ràng hơn.
 - + Lớp Low: Độ chính xác: 0.42, độ nhạy: 0.41 → Mặc dù có sự cải thiện so với các mô hình trước, nhưng lớp "Low" vẫn gặp khó khăn trong việc phân loại chính xác và cần sự cải thiện đáng kể.
 - + Lớp Standard: Độ chính xác: 0.86, độ nhạy: 0.85 → Lớp "Standard" tiếp tục được dự đoán chính xác nhất, với độ chính xác và độ nhạy cao, cho thấy mô hình hoạt động rất tốt đối với lớp này.
- Ma trận nhầm lẫn (Confusion Matrix):
 - + Lớp Standard: Dự đoán đúng 621 mẫu. Tuy nhiên, vẫn có 93 mẫu bị nhầm thành lớp "High" và 17 mẫu nhầm thành lớp "Low".
 - + Lớp High: Dự đoán đúng 257 mẫu. Tồn tại 93 mẫu bị nhầm thành lớp "Standard".
 - + Lớp Low: Chỉ có 15 mẫu được dự đoán đúng. 22 mẫu bị nhầm thành các lớp khác, chủ yếu là lớp "Standard".
- Tổng quan hiệu suất:
 - + Độ chính xác toàn cục (Accuracy): 0.78. Mô hình có độ chính xác khá cao (78%), cho thấy sự cải thiện trong việc phân loại các lớp so với các mô hình trước.
 - + Trung bình macro (Macro Average): Precision: 0.61, Recall: 0.61, F1-Score: 0.61 → Hiệu suất trung bình trên tất cả các lớp khá đồng đều, cho thấy mô hình có sự cân bằng trong việc phân loại các lớp mặc dù vẫn có sự nhầm lẫn với lớp "Low".

- + Trung bình trọng số (Weighted Average): Precision: 0.78, Recall: 0.78, F1-Score: 0.78
→ Do lớp "Standard" chiếm số lượng lớn, trung bình trọng số cao phản ánh mô hình hoạt động tốt trên lớp này.
- Nhận xét:
 - + Hiệu suất tốt nhất: Lớp "Standard" có hiệu suất cao nhất, với độ chính xác và độ nhạy vượt trội, phản ánh khả năng phân loại tốt của mô hình đối với lớp này.
 - + Nhược điểm lớn nhất: Lớp "Low" vẫn gặp khó khăn trong việc phân loại chính xác. Mặc dù có sự cải thiện nhẹ, nhưng lớp này vẫn bị nhầm lẫn nhiều và cần sự cải thiện rõ rệt.
 - + Mô hình 80/20 có sự cải thiện rõ rệt so với các mô hình trước, với độ chính xác cao đạt 78%. Lớp "Standard" vẫn được phân loại tốt nhất, trong khi lớp "Low" cần thêm các cải tiến để đạt được độ chính xác cao hơn. Cải thiện phân loại lớp "Low" và tiếp tục tối ưu hóa mô hình sẽ giúp nâng cao hiệu suất tổng thể.



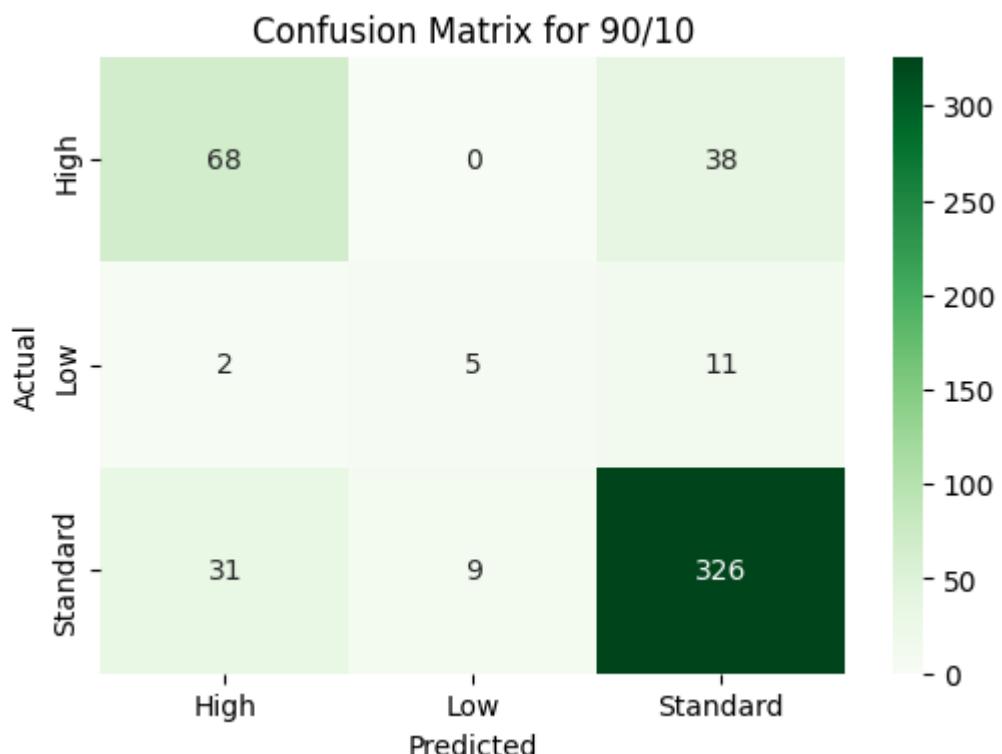
Classification Report cho tỷ lệ 90/10

- Hiệu suất của các lớp:
 - + Lớp High: Độ chính xác (Precision): 0.60, độ nhạy (Recall): 0.64 → Mô hình đã có sự cải thiện trong việc phân loại lớp "High", với độ nhạy tăng lên so với các mô hình trước. Tuy nhiên, vẫn còn nhầm lẫn trong việc phân loại lớp này.
 - + Lớp Low: Độ chính xác: 0.39, độ nhạy: 0.39 → Mặc dù mô hình đã cải thiện đôi chút, nhưng hiệu suất của lớp "Low" vẫn rất thấp và bị nhầm lẫn nhiều, cho thấy mô hình vẫn gặp khó khăn trong việc phân loại chính xác lớp này.
 - + Lớp Standard: Độ chính xác: 0.87, độ nhạy: 0.86 → Lớp "Standard" tiếp tục được dự đoán chính xác tốt, với độ chính xác và độ nhạy cao, phản ánh hiệu suất tốt của mô hình đối với lớp này.
- Ma trận nhầm lẫn (Confusion Matrix):
 - + Lớp High: Dự đoán đúng 68 mẫu. 38 mẫu bị nhầm thành các lớp khác.
 - + Lớp Low: Chỉ có 7 mẫu được dự đoán đúng. 11 mẫu bị nhầm thành các lớp khác, chủ yếu là lớp "Standard".
 - + Lớp Standard: Dự đoán đúng 312 mẫu. Một số mẫu bị nhầm thành lớp "High" và "Low", hưng phần lớn dự đoán chính xác.

	precision	recall	f1-score	support
High	0.6	0.64	0.62	106
Low	0.39	0.39	0.39	18
Standard	0.87	0.85	0.86	366
accuracy			0.79	490
macro average	0.62	0.63	0.62	490
weighted average	0.79	0.79	0.79	490

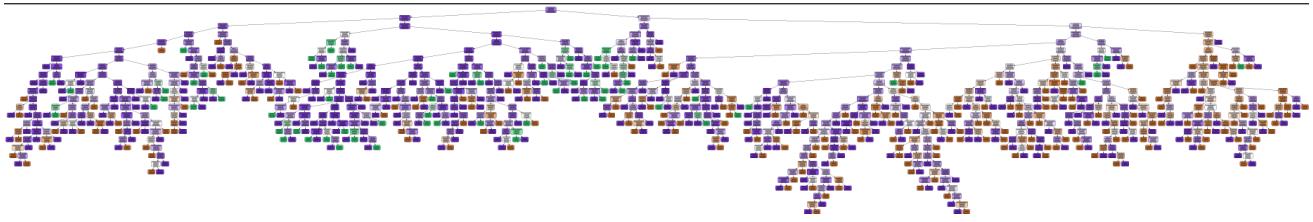
- Tổng quan hiệu suất:
 - + Độ chính xác toàn cục (Accuracy): 0.79. Mô hình đạt độ chính xác cao (79%), cho thấy khả năng phân loại tổng thể khá tốt, nhưng vẫn có một số vấn đề với lớp "Low".
 - + Trung bình macro (Macro Average): Precision: 0.62, Recall: 0.63, F1-Score: 0.62 → Hiệu suất trung bình trên tất cả các lớp cho thấy mô hình đang phân loại tương đối đồng đều, nhưng vẫn cần cải thiện, đặc biệt là đối với lớp "Low".

- + Trung bình trọng số (Weighted Average): Precision: 0.79, Recall: 0.79, F1-Score: 0.79
→ Trung bình trọng số phản ánh hiệu suất tốt của mô hình đối với lớp "Standard", chiếm ưu thế về số lượng, giúp đạt được giá trị cao trong tất cả các chỉ số
- Nhận xét:
 - + Hiệu suất tốt nhất: Lớp "Standard" có hiệu suất tốt nhất, với độ chính xác và độ nhạy cao, phản ánh khả năng phân loại chính xác lớp này trong mô hình
 - + Nhược điểm lớn nhất: Lớp "Low" vẫn gặp khó khăn lớn trong việc phân loại chính xác. Mặc dù đã có cải thiện nhẹ, nhưng hiệu suất của lớp này vẫn rất thấp, đặc biệt là trong việc tránh bị nhầm lẫn với lớp "Standard".
 - + Mô hình 90/10 đã đạt được độ chính xác 79%, một cải thiện so với các mô hình trước, đặc biệt là đối với lớp "High". Tuy nhiên, lớp "Low" vẫn gặp khó khăn trong việc phân loại chính xác, và cần có những cải tiến đáng kể, đặc biệt là trong việc giảm nhầm lẫn với lớp "Standard". Nếu mô hình được tối ưu hóa thêm cho lớp "Low", hiệu suất tổng thể sẽ được cải thiện rõ rệt.

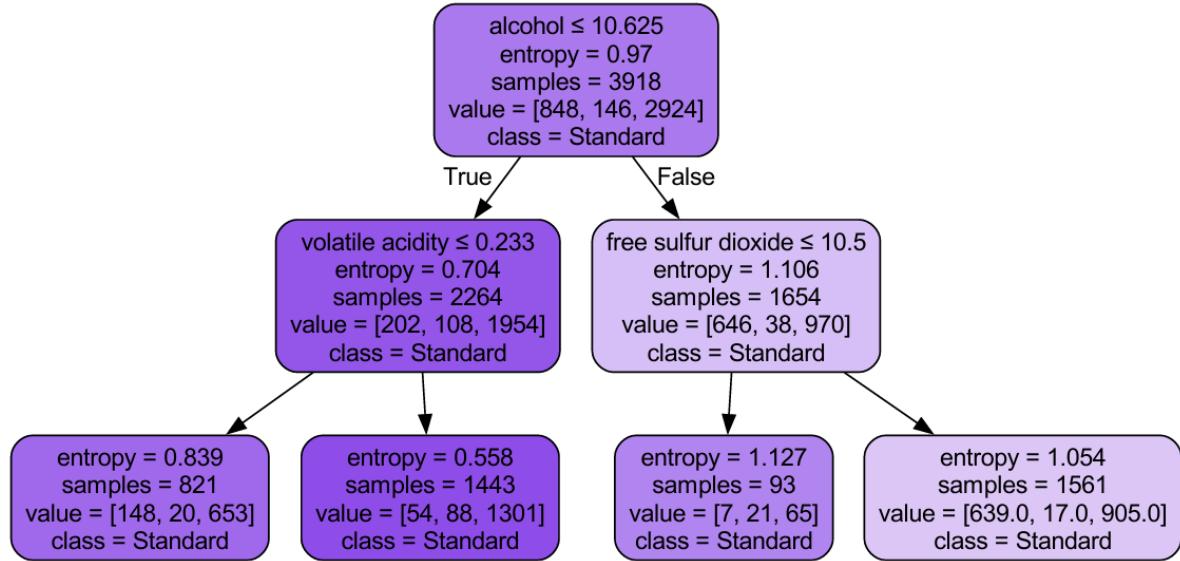


2.5. Ảnh hưởng của độ sâu đối với độ chính xác của cây quyết định

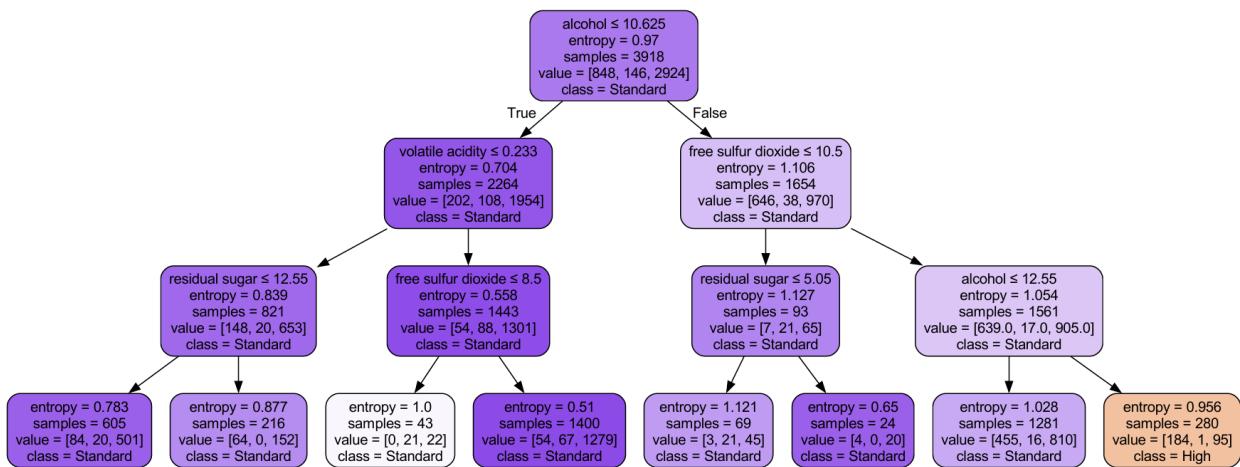
Ảnh cây quyết định với max_depth = None: Results/TreeVisualizationsWithMaxDepth/none.pdf



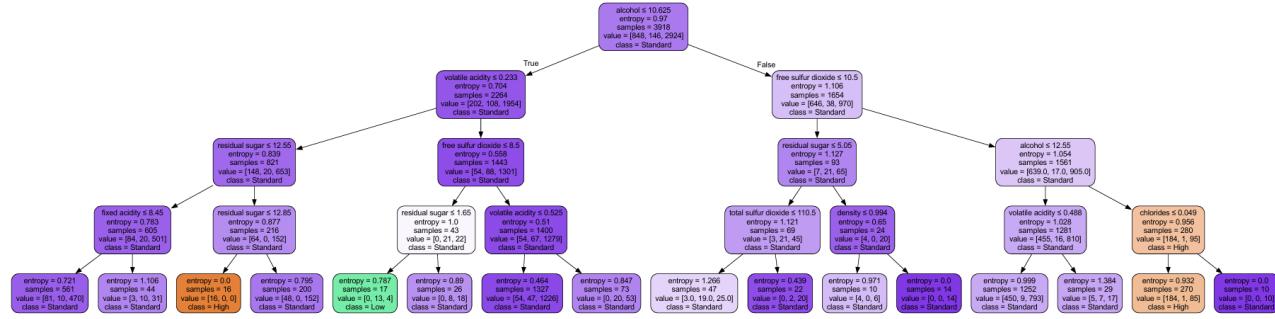
Ảnh cây quyết định với max_depth = 2: Results/TreeVisualizationsWithMaxDepth/2.pdf



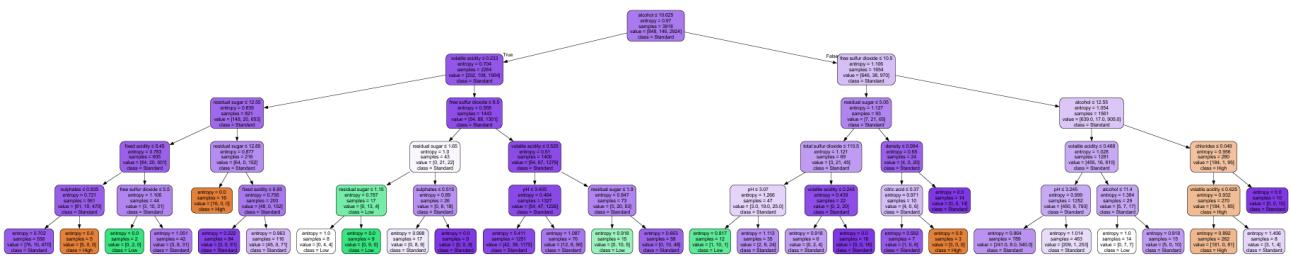
Ảnh cây quyết định với max_depth = 3: Results/TreeVisualizationsWithMaxDepth/3.pdf



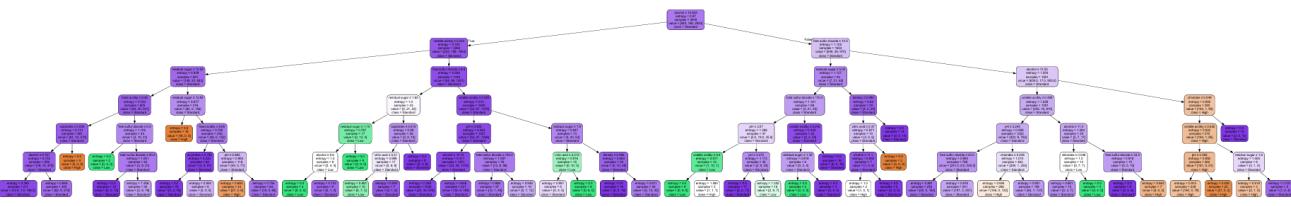
Ảnh cây quyết định với max_depth = 4: Results/TreeVisualizationsWithMaxDepth/4.pdf



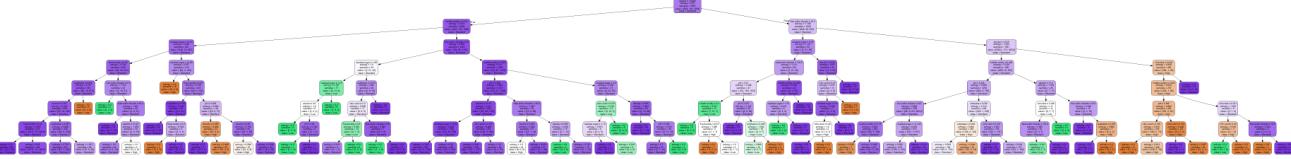
Ảnh cây quyết định với max_depth = 5: Results/TreeVisualizationsWithMaxDepth/5.pdf



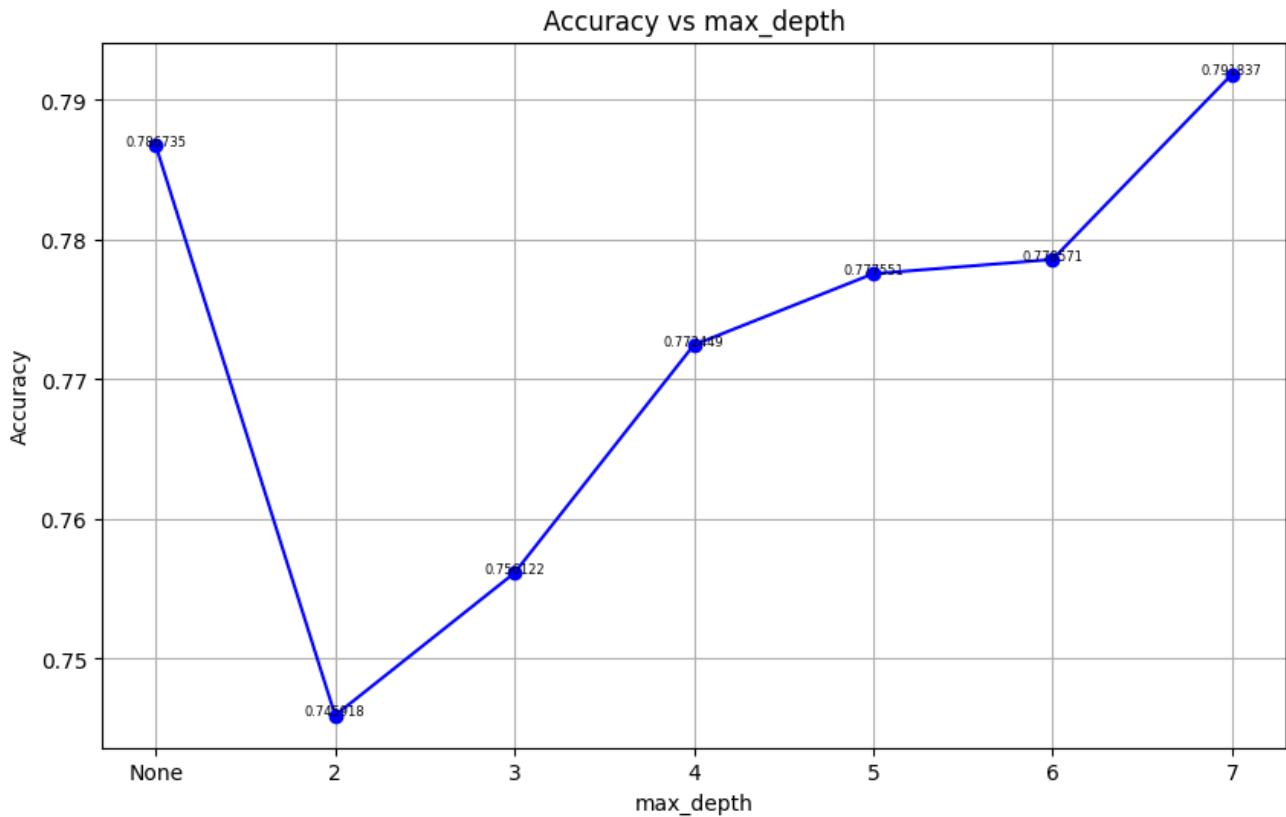
Ảnh cây quyết định với max_depth = 6: Results/TreeVisualizationsWithMaxDepth/6.pdf



Ảnh cây quyết định với max_depth = 7: Results/TreeVisualizationsWithMaxDepth/7.pdf



max_depth	None	2	3	4	5	6	7
Accuracy	0.7312	0.7660	0.7820	0.7857	0.7801	0.7579	0.7547



- Xu hướng ban đầu: Khi giá trị `max_depth` tăng từ 2 lên 7, độ chính xác của mô hình dần cải thiện và ổn định. Mặc dù `max_depth = None` đạt độ chính xác khá cao (~77.5%), nhưng không vượt trội so với các giá trị giới hạn như 6 hoặc 7.
- Ý nghĩa của `max_depth = None`: Khi không giới hạn độ sâu (`max_depth = None`), độ chính xác chỉ đạt khoảng 77.5%. Điều này cho thấy việc không giới hạn độ sâu cho phép mô hình học quá nhiều thông tin từ dữ liệu, nhưng kết quả không vượt trội hơn so với các giá trị giới hạn nhỏ hơn như 6 hoặc 7.
- Điểm thấp nhất: Với `max_depth = 2`, độ chính xác thấp nhất (~74.5%) cho thấy mô hình quá đơn giản và không đủ khả năng học được cấu trúc phức tạp của dữ liệu.
- Xu hướng cải thiện: Độ chính xác tăng dần từ `max_depth = 3` đến `max_depth = 7`. Cả hai giá trị `max_depth = 6` và `max_depth = 7` đạt độ chính xác cao nhất (~78.4%), nhưng không có sự cải thiện đáng kể khi tăng thêm độ sâu sau giá trị này.
- Kết luận: Giá trị `max_depth = 6` hoặc `max_depth = 7` là tối ưu nhất, vì chúng mang lại độ chính xác cao nhất trong các thử nghiệm. Không có dấu hiệu rõ ràng của hiện tượng overfitting trong khoảng giá trị đã thử nghiệm. Việc không giới hạn độ sâu (với `max_depth = None`) không mang lại lợi ích rõ rệt so với việc đặt giới hạn cụ thể như 6 hoặc 7.

3. Breast Cancer dataset

3.1. Tập dữ liệu

Nguồn: [Breast Cancer Wisconsin \(Diagnostic\) - UCI Machine Learning Repository](#)

Thông tin:

- Số mẫu: 569
- Số đặc trưng: 30 (các thông tin về hình ảnh)
- Mục tiêu: Phân loại khối u là lành tính (B) hoặc ác tính (M).

3.2. Chuẩn bị dữ liệu

Đọc dữ liệu từ tệp CSV bằng thư viện pandas.

Kiểm tra dữ liệu:

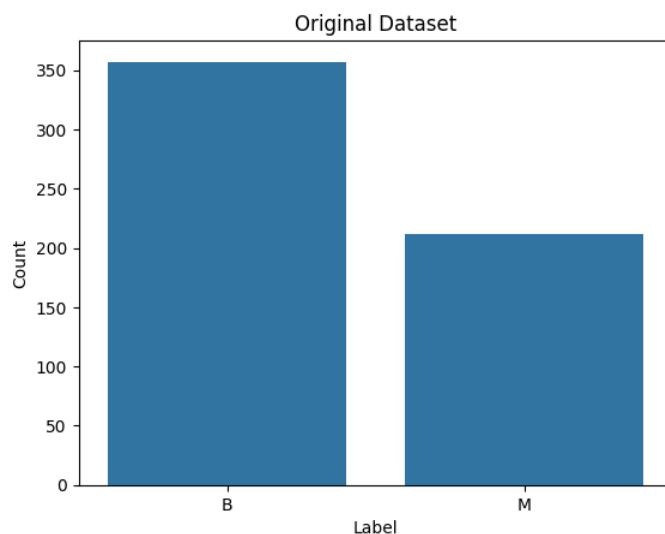
- Xác định số mẫu (569), số đặc trưng (30), và nhãn (Benign/Malignant).
- Kiểm tra giá trị thiếu và thay thế (nếu có).

Chia dữ liệu thành các tập:

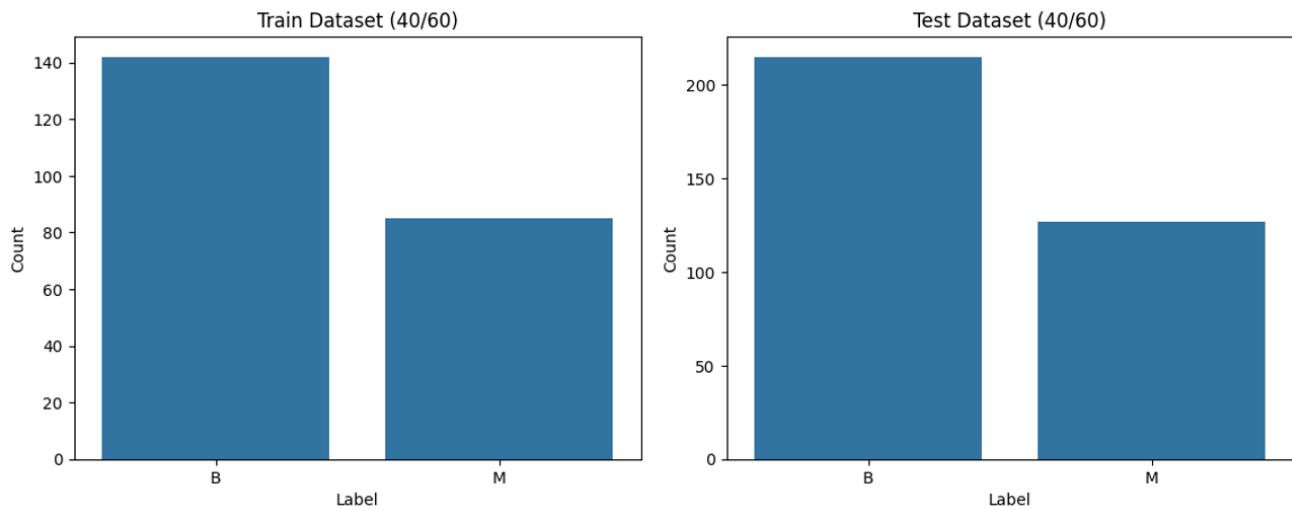
- Tập train/test theo các tỷ lệ 40/60, 60/40, 80/20, 90/10.
- Sử dụng phương pháp StratifiedShuffleSplit để giữ nguyên phân phối nhãn.

Trực quan hóa: Vẽ biểu đồ cột (bar chart) phân phối nhãn trong tập train/test.

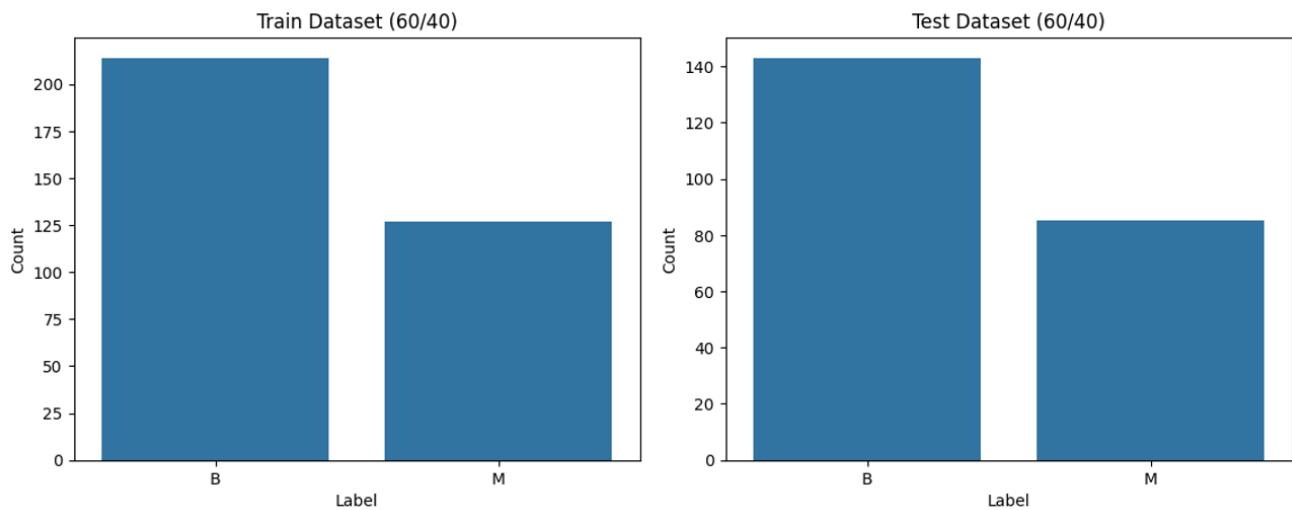
Dữ liệu gốc



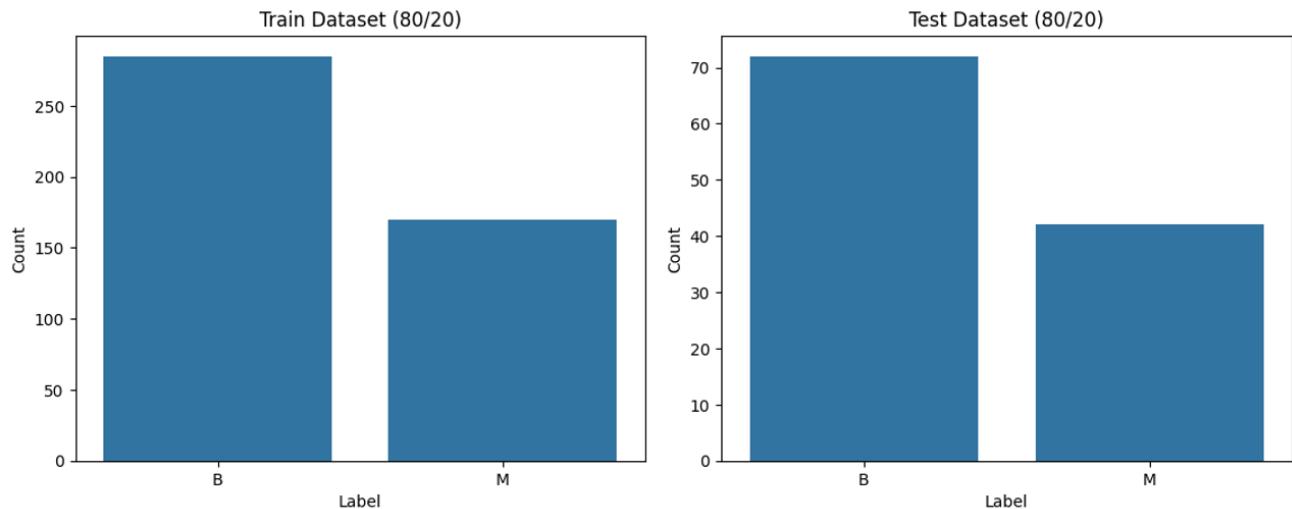
Tập train/test theo các tỷ lệ 40/60



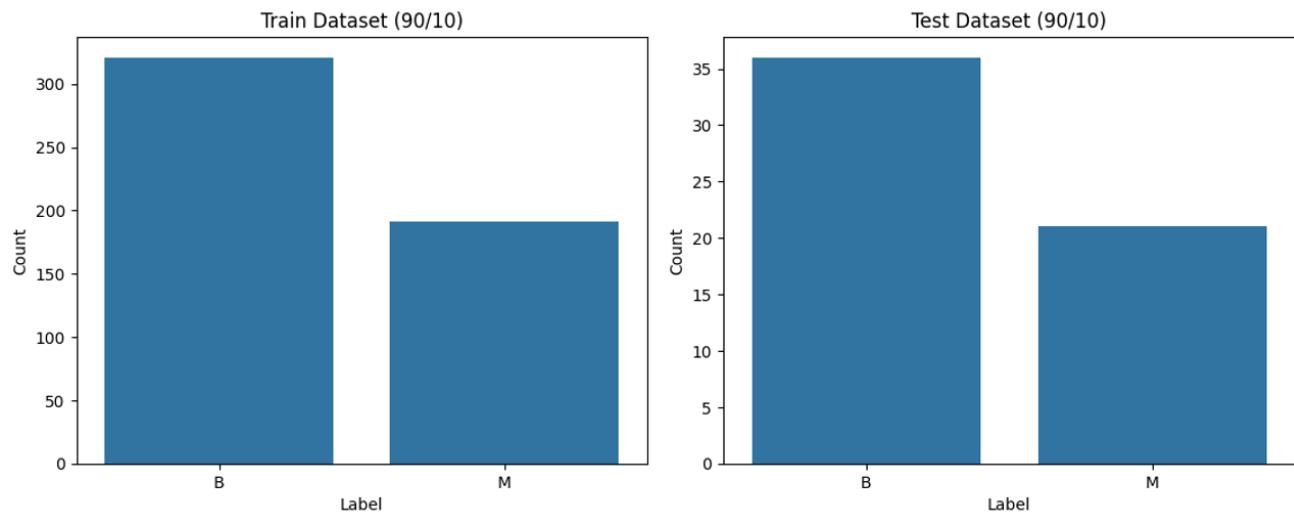
Tập train/test theo các tỷ lệ 60/40



Tập train/test theo các tỷ lệ 80/20



Tập train/test theo các tỷ lệ 90/10



3.3. Xây dựng mô hình cây quyết định

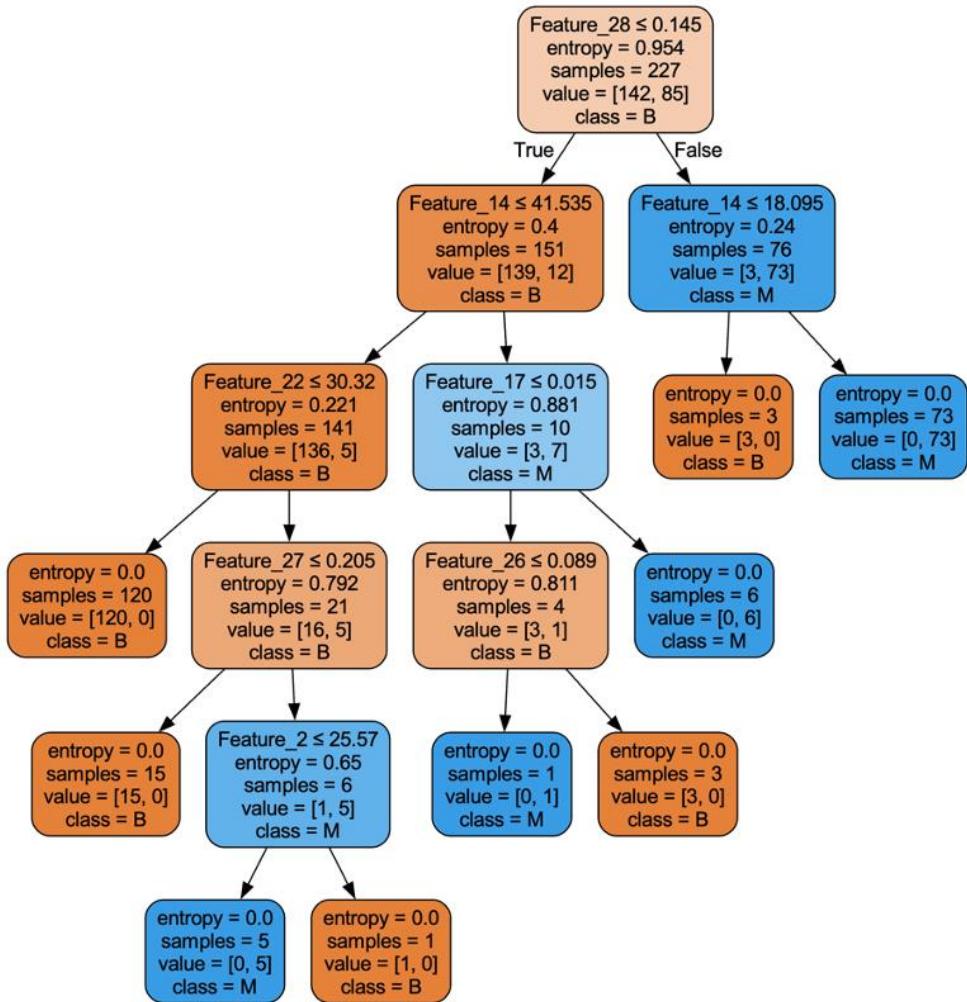
Công cụ sử dụng: DecisionTreeClassifier của scikit-learn.

Thông số quan trọng:

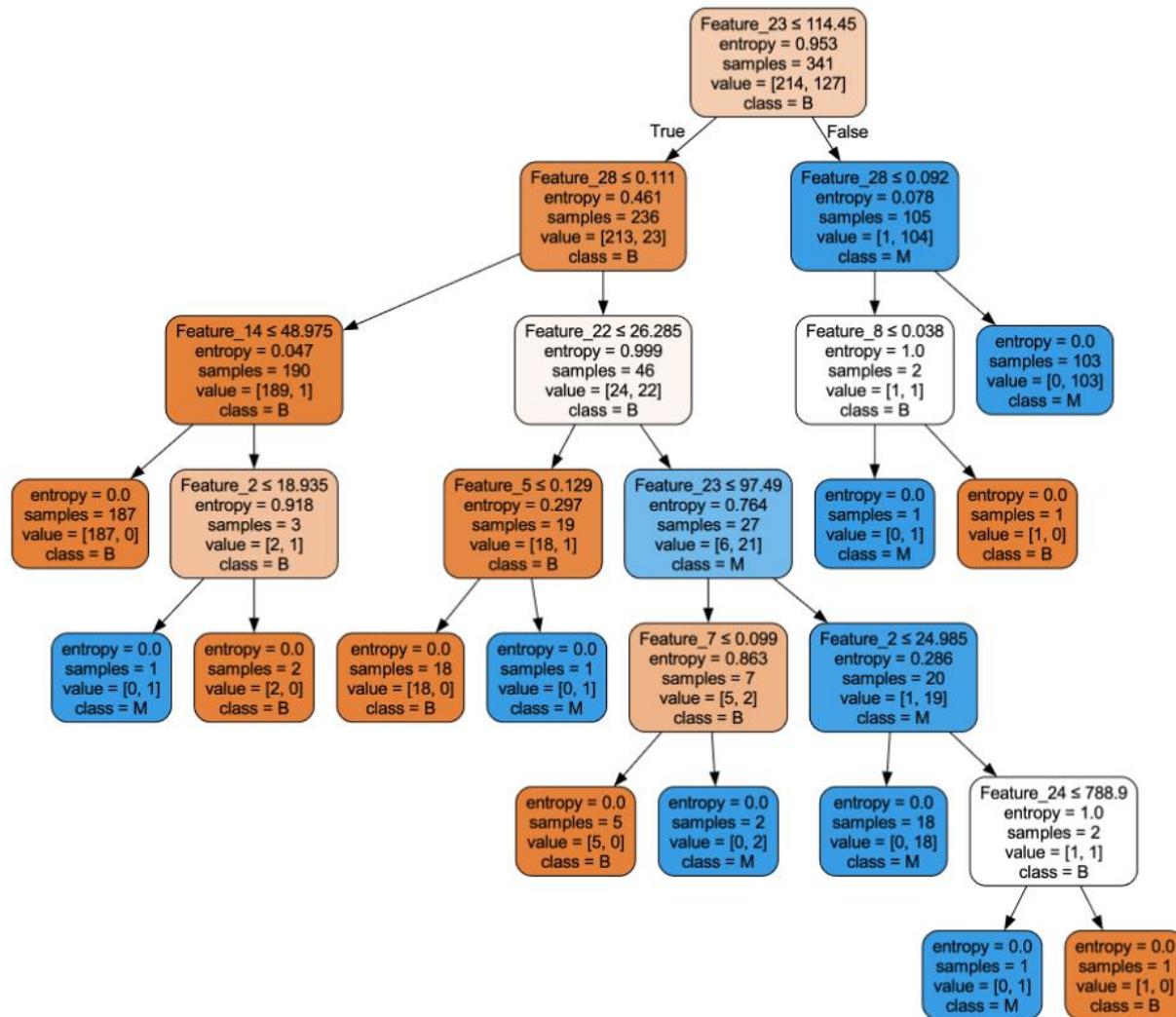
- Tiêu chí phân tách: Information Gain (entropy).
- Hiển thị cây bằng Graphviz.

Kết quả mong đợi: Cây quyết định được hiển thị tương ứng với từng tỷ lệ train/test.

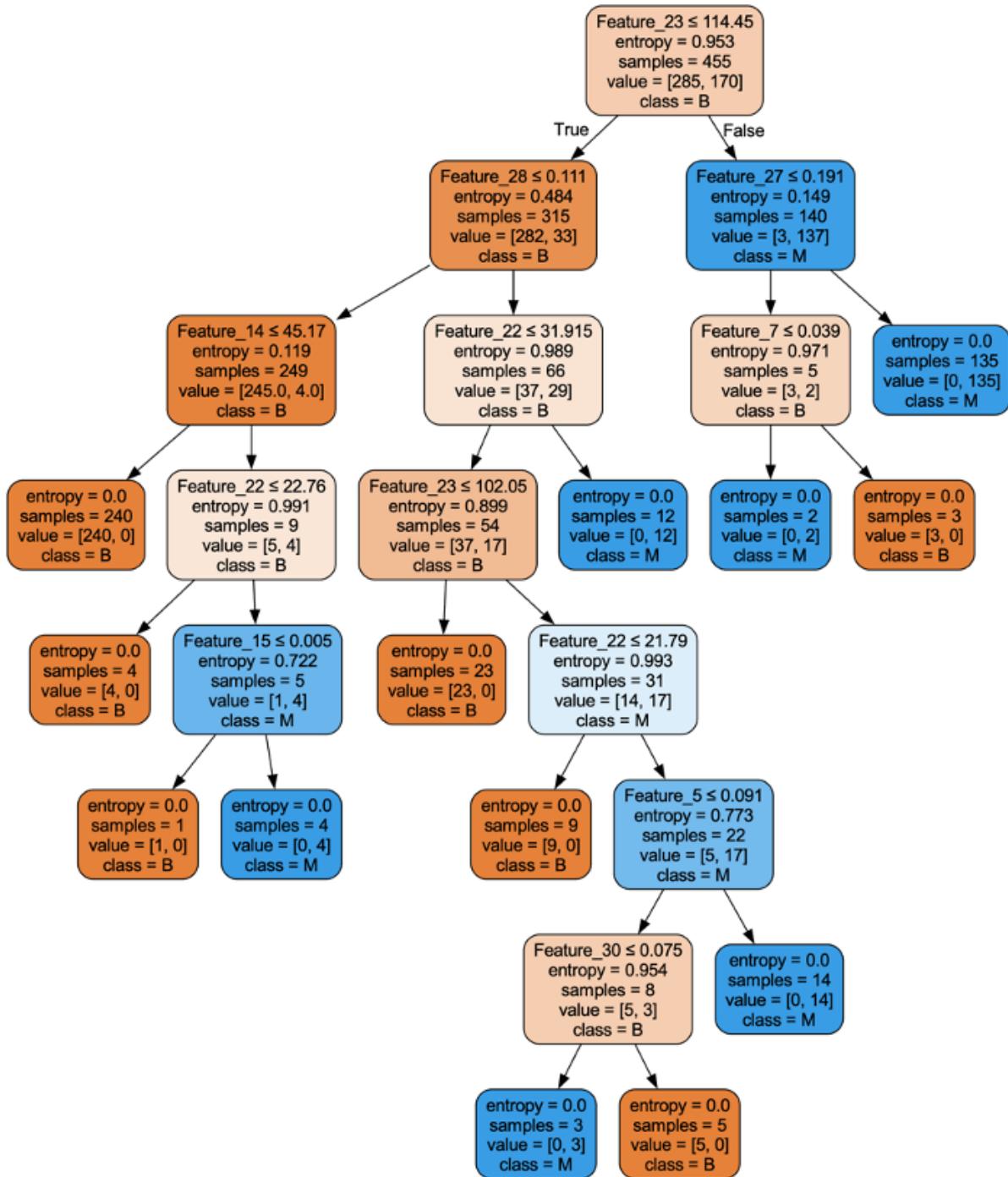
Tỉ lệ tập train/test là 40/60



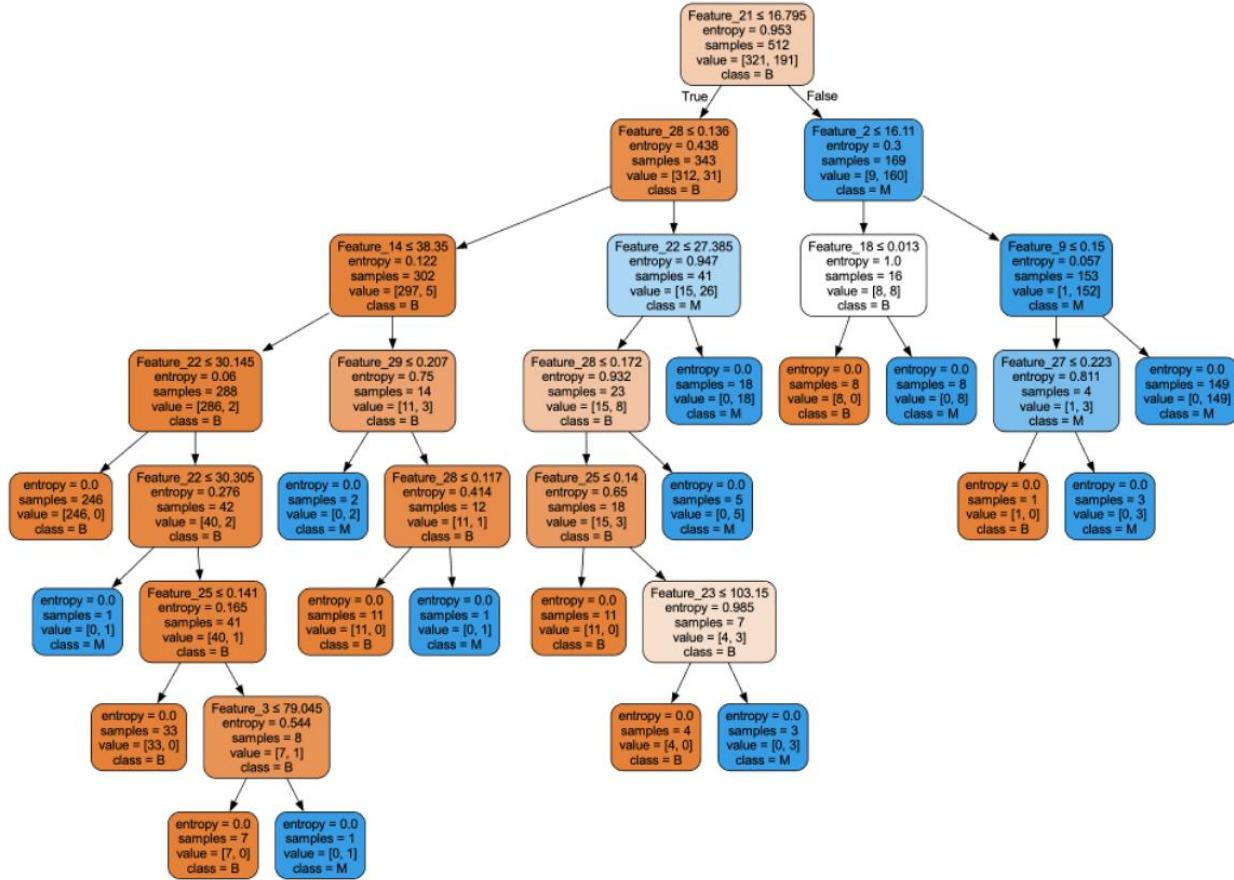
Tỉ lệ tập train/test là 60/40



Tỉ lệ tập train/test là 80/20



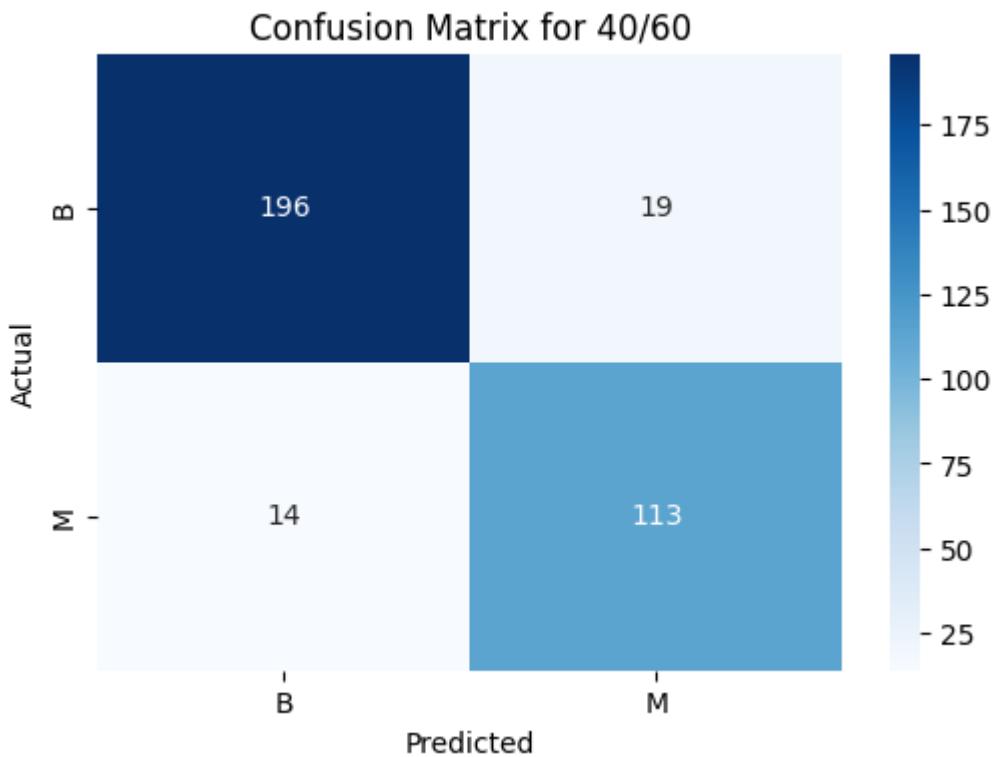
Tỉ lệ tập train/test là 90/10



3.4. Đánh giá mô hình cây quyết định

	precision	recall	f1-score	support
B	0.93	0.91	0.92	215
M	0.86	0.89	0.87	127
accuracy			0.90	342
macro avg	0.89	0.90	0.90	342
weighted avg	0.90	0.90	0.90	342

Tỉ lệ tập train/test là 40/60



Confusion Matrix:

- True Positive (TP): 113 (Mô hình dự đoán đúng nhãn 1 (M)).
- True Negative (TN): 196 (Mô hình dự đoán đúng nhãn 0 (B)).
- False Positive (FP): 19 (Mô hình dự đoán nhãn 1 (M) nhưng thực tế là 0 (B)).
- False Negative (FN): 14 (Mô hình dự đoán nhãn 0 (B) nhưng thực tế là 1 (M)).

Classification Report:

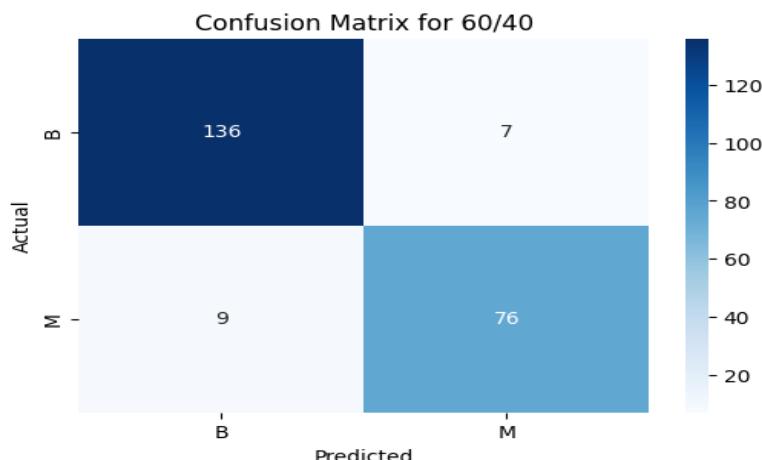
- Lớp B (Benign):
 - + Precision: 0.93 (Khả năng dự đoán đúng các mẫu B là rất cao).
 - + Recall: 0.91 (Mô hình phát hiện được phần lớn các mẫu thực sự là B).
 - + F1-Score: 0.92 (Cân bằng giữa precision và recall, rất tốt).
- Lớp M (Malignant):
 - + Precision: 0.86 (Một số mẫu được dự đoán là M nhưng sai).
 - + Recall: 0.89 (Mô hình phát hiện hầu hết các mẫu thực sự là M).
 - + F1-Score: 0.87 (Hiệu suất ở lớp này hơi thấp hơn lớp B).
- Accuracy: Với giá trị 90%, mô hình hoạt động khá tốt. Tuy nhiên, độ chính xác không đủ để đánh giá toàn diện nếu dữ liệu bị mất cân bằng.
- Macro Average: Giá trị (0.89 - 0.90) cho thấy mô hình hoạt động đồng đều ở cả hai lớp.
- Weighted Average: Giá trị (0.90) thể hiện mô hình được tối ưu hóa tốt trên tổng thể dữ liệu.

Dánh giá tổng thể cây quyết định:

- Mô hình cây quyết định đạt hiệu suất tốt, với độ chính xác cao (90%) và các chỉ số khác cân bằng.
- Có thể cải thiện precision ở lớp M để giảm các dự đoán sai
- Tập trung vào việc giảm nhầm lẫn từ B sang M và ngược lại.

	precision	recall	f1-score	support
B	0.94	0.95	0.94	143
M	0.92	0.89	0.90	85
accuracy			0.93	228
macro avg	0.93	0.92	0.92	228
weighted avg	0.93	0.93	0.93	228

Tỉ lệ tập train/test là 60/40



Confusion Matrix:

- True Positive (TP): 76 (Mô hình dự đoán đúng nhãn 1 (M)).
- True Negative (TN): 136 (Mô hình dự đoán đúng nhãn 0 (B)).
- False Positive (FP): 7 (Mô hình dự đoán nhãn 1 (M) nhưng thực tế là 0 (B)).
- False Negative (FN): 9 (Mô hình dự đoán nhãn 0 (B) nhưng thực tế là 1 (M)).

Classification Report:

- Lớp B (Benign):
 - + Precision: 0.94 (Khả năng dự đoán đúng các mẫu B là rất cao).

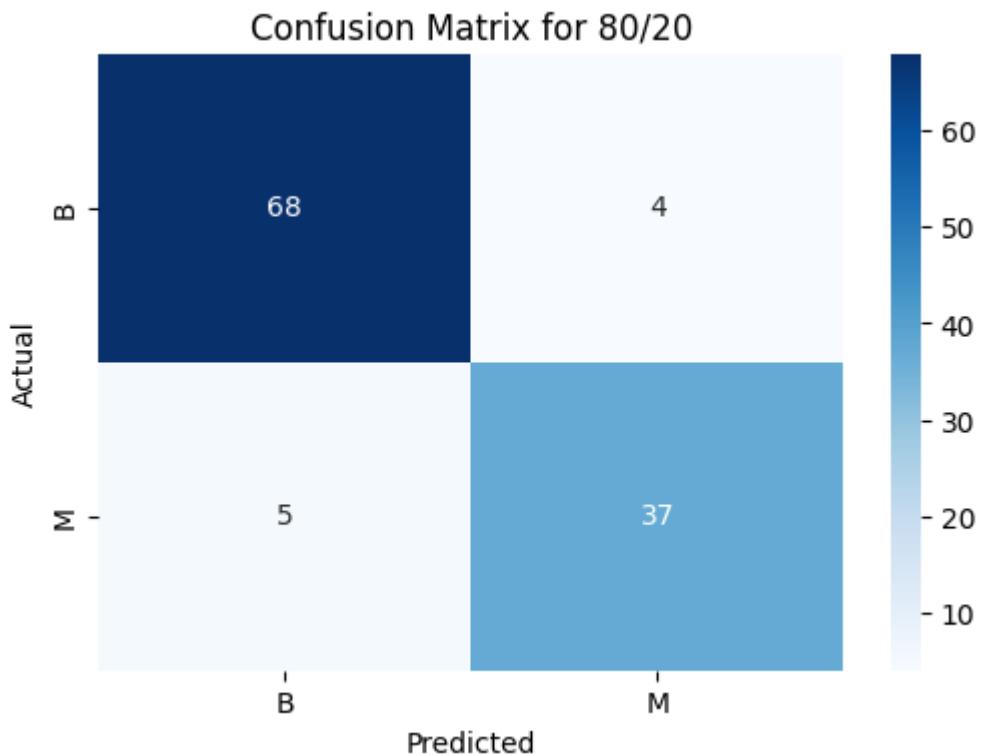
- + Recall: 0.95 (Mô hình phát hiện được phần lớn các mẫu thực sự là B).
- + F1-Score: 0.94 (Cân bằng giữa precision và recall, rất tốt).
- Lớp M (Malignant):
 - + Precision: 0.92 (Một số mẫu được dự đoán là M nhưng sai).
 - + Recall: 0.89 (Mô hình phát hiện hầu hết các mẫu thực sự là M).
 - + F1-Score: 0.90 (Hiệu suất ở lớp này hơi thấp hơn lớp B).
- Accuracy: Với giá trị 93%, mô hình hoạt động tốt hơn so với tỷ lệ 40/60. Độ chính xác cao thể hiện mô hình có khả năng phân loại tốt trên tổng thể dữ liệu, ngay cả khi có sự cân bằng giữa hai lớp.
- Macro Average: Giá trị (0.92 - 0.93) cho thấy mô hình hoạt động đồng đều ở cả hai lớp.
- Weighted Average: Giá trị (0.93) thể hiện mô hình được tối ưu hóa tốt trên tổng thể dữ liệu, với trọng số ưu tiên cho lớp có nhiều mẫu hơn.

Đánh giá tổng thể cây quyết định:

- Mô hình cây quyết định đạt hiệu suất tốt, với độ chính xác cao (93%) và các chỉ số khác cân bằng.
- Cần cải thiện recall ở lớp M để phát hiện thêm các mẫu ác tính.
- Tập trung vào việc giảm nhầm lẫn từ M sang B (9 mẫu) và từ B sang M (7 mẫu) để nâng cao độ chính xác hơn nữa.
- Tỷ lệ train/test 60/40 cho thấy mô hình học tốt hơn nhờ tập huấn luyện lớn hơn.

	precision	recall	f1-score	support
B	0.93	0.94	0.94	72
M	0.90	0.88	0.89	42
accuracy			0.92	114
macro avg	0.92	0.91	0.91	114
weighted avg	0.92	0.92	0.92	114

Tỉ lệ tập train/test là 80/20



Confusion Matrix:

- True Positive (TP): 37 (Mô hình dự đoán đúng nhãn 1 (M)).
- True Negative (TN): 68 (Mô hình dự đoán đúng nhãn 0 (B)).
- False Positive (FP): 4 (Mô hình dự đoán nhãn 1 (M) nhưng thực tế là 0 (B)).
- False Negative (FN): 5 (Mô hình dự đoán nhãn 0 (B) nhưng thực tế là 1 (M)).

Classification Report:

- Lớp B (Benign):
 - + Precision: 0.93 (Khả năng dự đoán đúng các mẫu B là rất cao).
 - + Recall: 0.94 (Mô hình phát hiện được phần lớn các mẫu thực sự là B).
 - + F1-Score: 0.94 (Cân bằng giữa precision và recall, rất tốt).
- Lớp M (Malignant):
 - + Precision: 0.90 (Một số mẫu được dự đoán là M nhưng sai).
 - + Recall: 0.88 (Mô hình phát hiện hầu hết các mẫu thực sự là M).
 - + F1-Score: 0.89 (Hiệu suất ở lớp này hơi thấp hơn lớp B).
- Accuracy: Với giá trị 92%, mô hình hoạt động khá tốt trên tổng thể. Độ chính xác cao nhưng cần xem xét kỹ chỉ số recall ở lớp M.
- Macro Average: Giá trị (0.91 - 0.92) cho thấy mô hình hoạt động đồng đều giữa hai lớp.

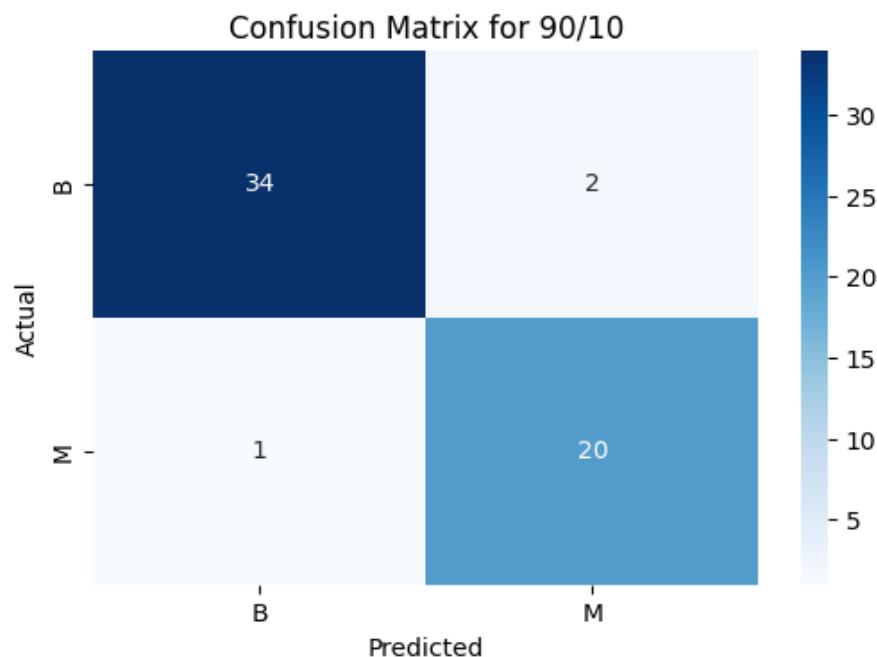
- Weighted Average: Giá trị (0.92) thể hiện mô hình được tối ưu hóa tốt trên toàn bộ dữ liệu, đặc biệt cân nhắc số lượng mẫu trong từng lớp.

Dánh giá tổng thể cây quyết định:

- Mô hình cây quyết định đạt hiệu suất tốt với độ chính xác cao (92%) và các chỉ số khác tương đối cân bằng.
- Cần cải thiện recall ở lớp M để phát hiện thêm các mẫu ác tính.
- Tập trung giảm nhầm lẫn từ M sang B (5 mẫu) và từ B sang M (4 mẫu).
- Tỷ lệ train/test 80/20 vẫn đảm bảo hiệu suất tốt, cho thấy mô hình đủ khả năng học với tập huấn luyện lớn hơn.

	precision	recall	f1-score	support
B	0.97	0.94	0.96	36
M	0.91	0.95	0.93	21
accuracy			0.95	57
macro avg	0.94	0.95	0.94	57
weighted avg	0.95	0.95	0.95	57

Tỉ lệ tập train/test là 90/10



Confusion Matrix:

- True Positive (TP): 20 (Mô hình dự đoán đúng nhãn 1 (M)).
- True Negative (TN): 34 (Mô hình dự đoán đúng nhãn 0 (B)).
- False Positive (FP): 1 (Mô hình dự đoán nhãn 1 (M) nhưng thực tế là 0 (B)).
- False Negative (FN): 2 (Mô hình dự đoán nhãn 0 (B) nhưng thực tế là 1 (M)).

Classification Report:

- Lớp B (Benign):
 - + Precision: 0.97 (Khả năng dự đoán đúng các mẫu B là rất cao).
 - + Recall: 0.94 (Mô hình phát hiện được phần lớn các mẫu thực sự là B).
 - + F1-Score: 0.96 (Cân bằng giữa precision và recall, rất tốt).
- Lớp M (Malignant):
 - + Precision: 0.91 (Một số mẫu được dự đoán là M nhưng sai).
 - + Recall: 0.95 (Mô hình phát hiện gần như tất cả các mẫu thực sự là M).
 - + F1-Score: 0.93 (Hiệu suất tổng thể ở lớp M là rất tốt).
- Accuracy: Với giá trị 95%, mô hình hoạt động rất tốt trên tổng thể, thể hiện khả năng dự đoán chính xác cao.
- Macro Average: Giá trị (0.94 - 0.95) cho thấy mô hình hoạt động đồng đều ở cả hai lớp.
- Weighted Average: Giá trị (0.95) thể hiện mô hình được tối ưu hóa tốt trên toàn bộ dữ liệu, với trọng số ưu tiên cho lớp có nhiều mẫu hơn.

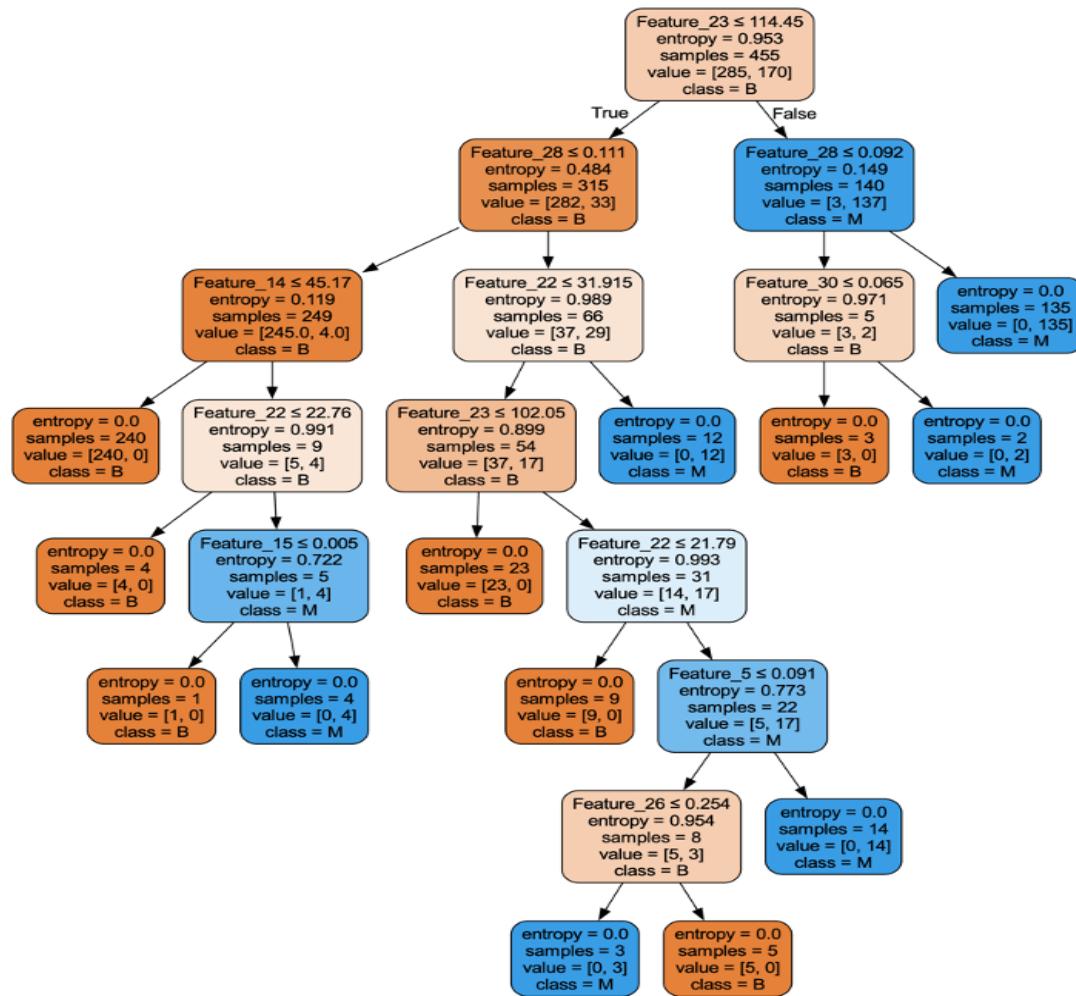
Đánh giá hiệu suất:

- Mô hình đạt hiệu suất rất tốt với độ chính xác cao (95%) và các chỉ số khác cân bằng.
- Tỷ lệ nhầm lẫn rất thấp (1 mẫu FP và 2 mẫu FN), chứng tỏ khả năng phân loại hiệu quả.
- Tuy nhiên, vẫn có thể cải thiện precision ở lớp M để giảm các dự đoán sai.
- Tỷ lệ train/test 90/10 cho thấy mô hình hoạt động hiệu quả ngay cả với tập kiểm tra nhỏ.

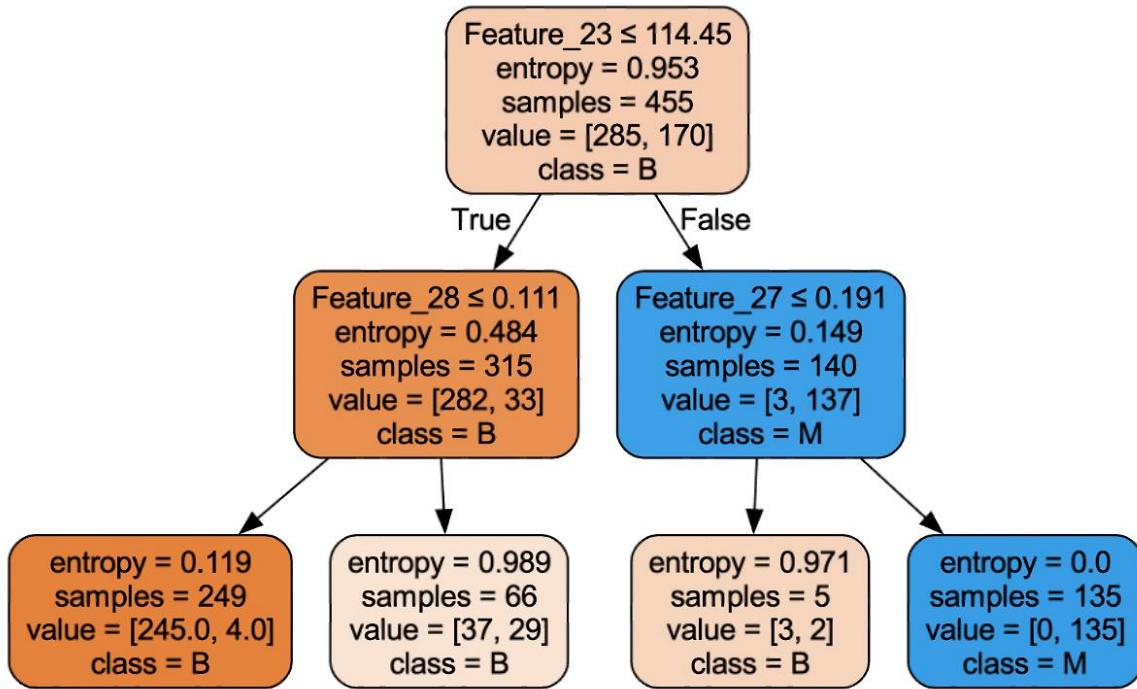
3.5. Ảnh hưởng của độ sâu đối với độ chính xác của cây quyết định

Biểu đồ mô hình cây ứng với từng max depth

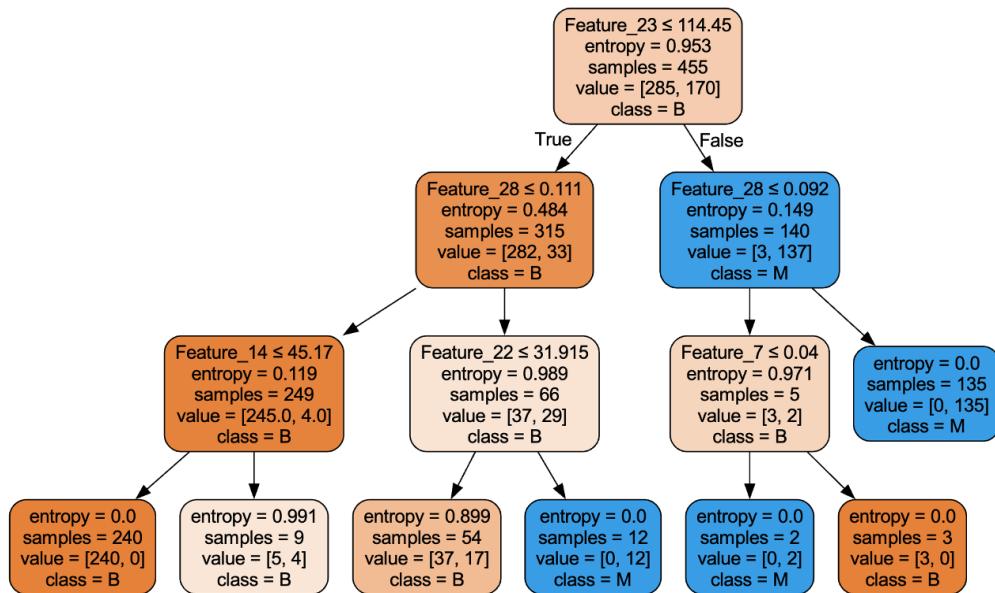
Ảnh cây quyết định với max_depth = None



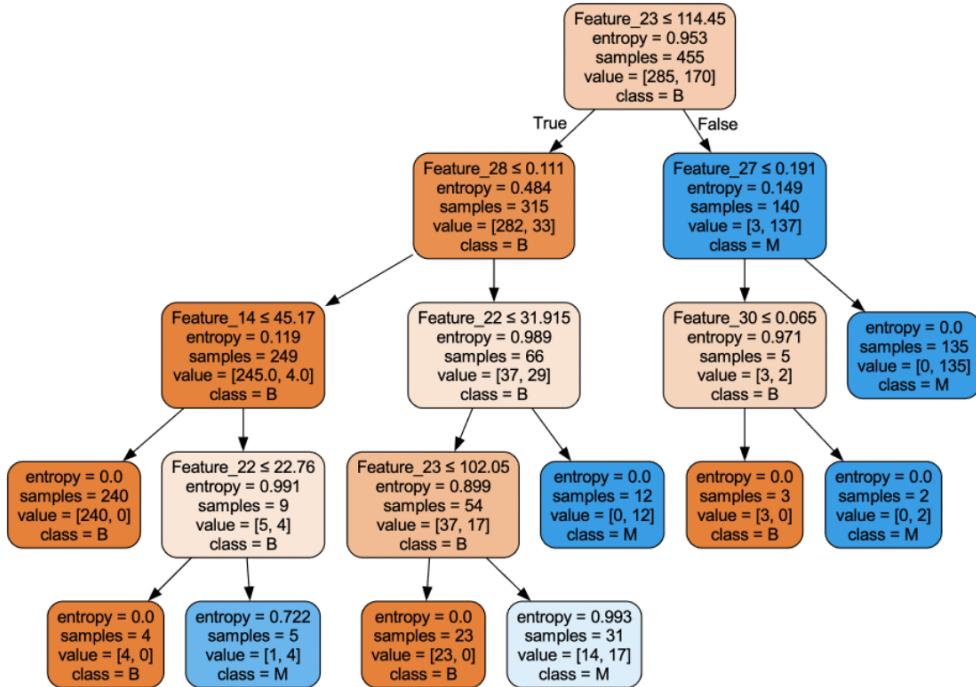
Ảnh cây quyết định với max_depth = 2



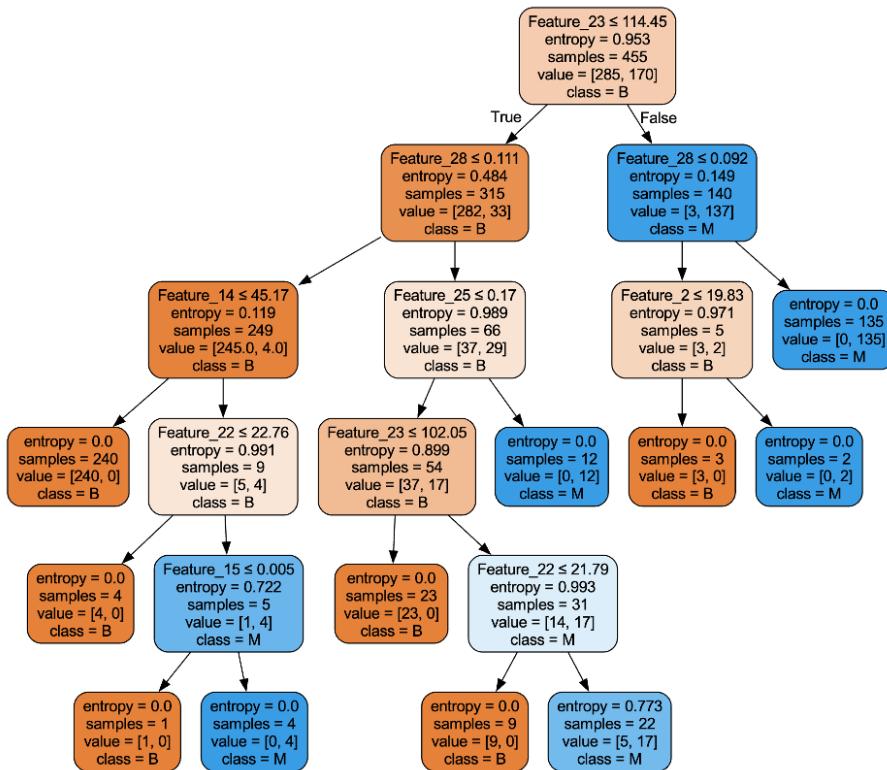
Ảnh cây quyết định với max_depth = 3



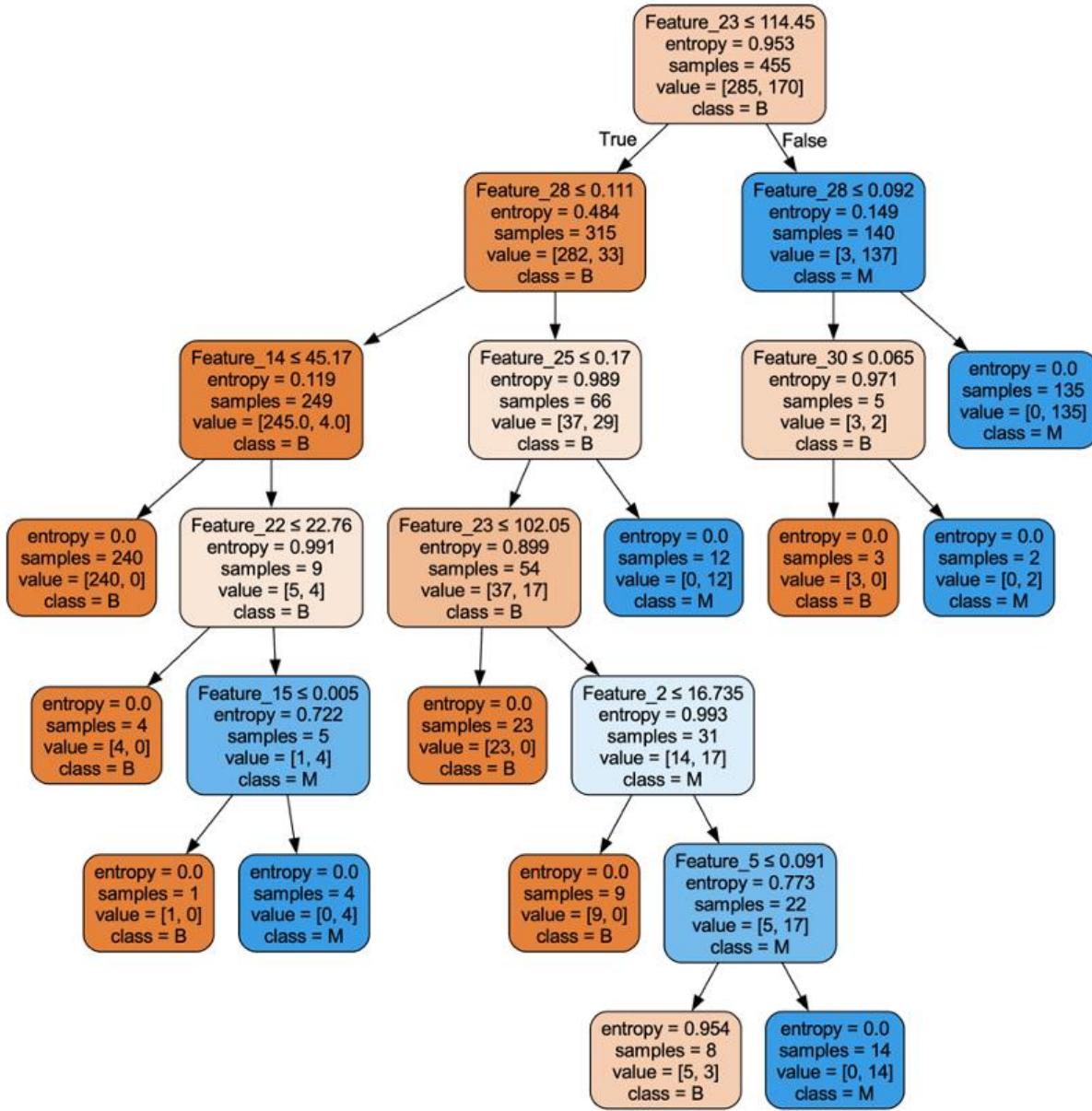
Ảnh cây quyết định với max_depth = 4



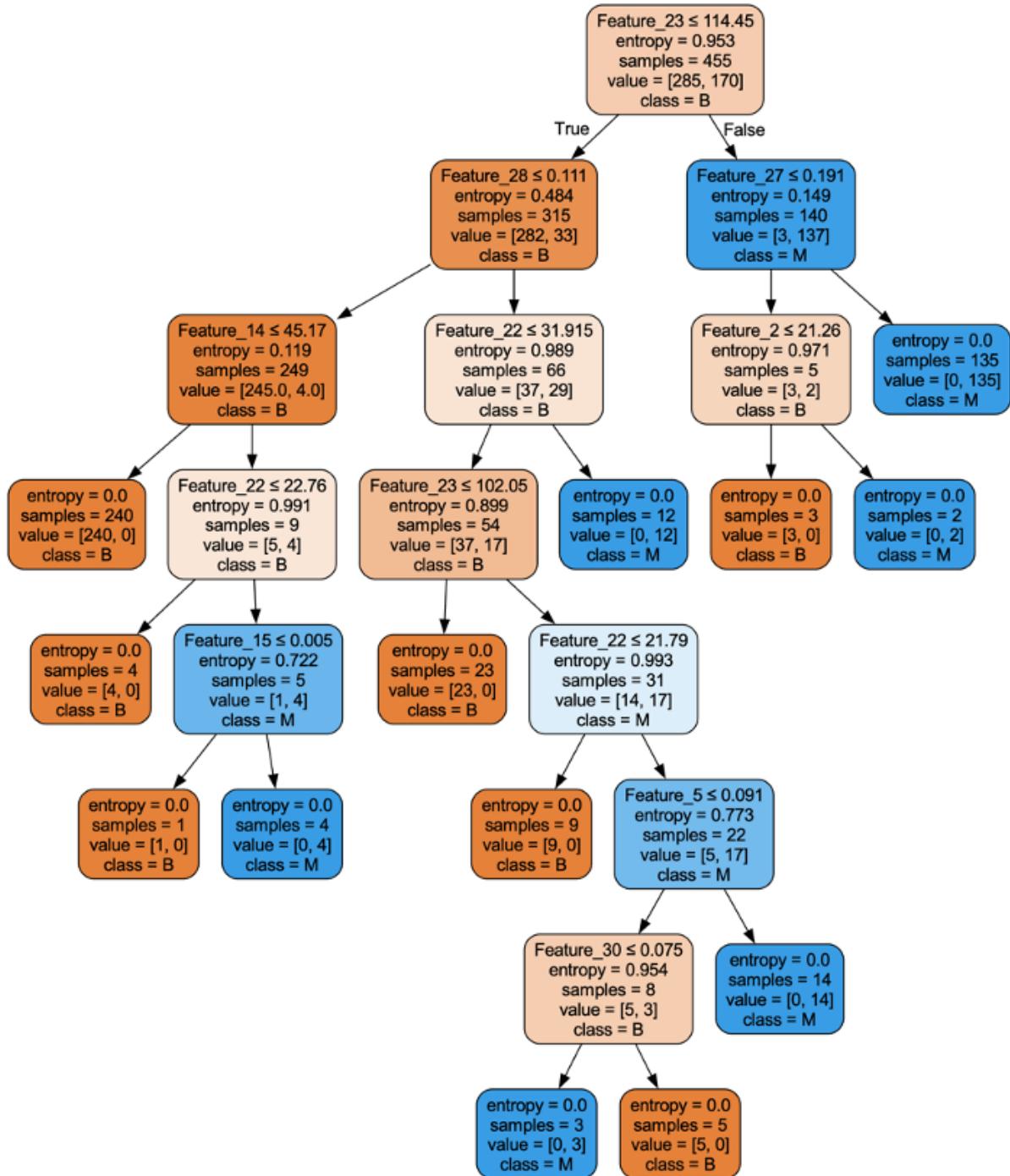
Ảnh cây quyết định với max_depth = 5



Ảnh cây quyết định với max_depth = 6

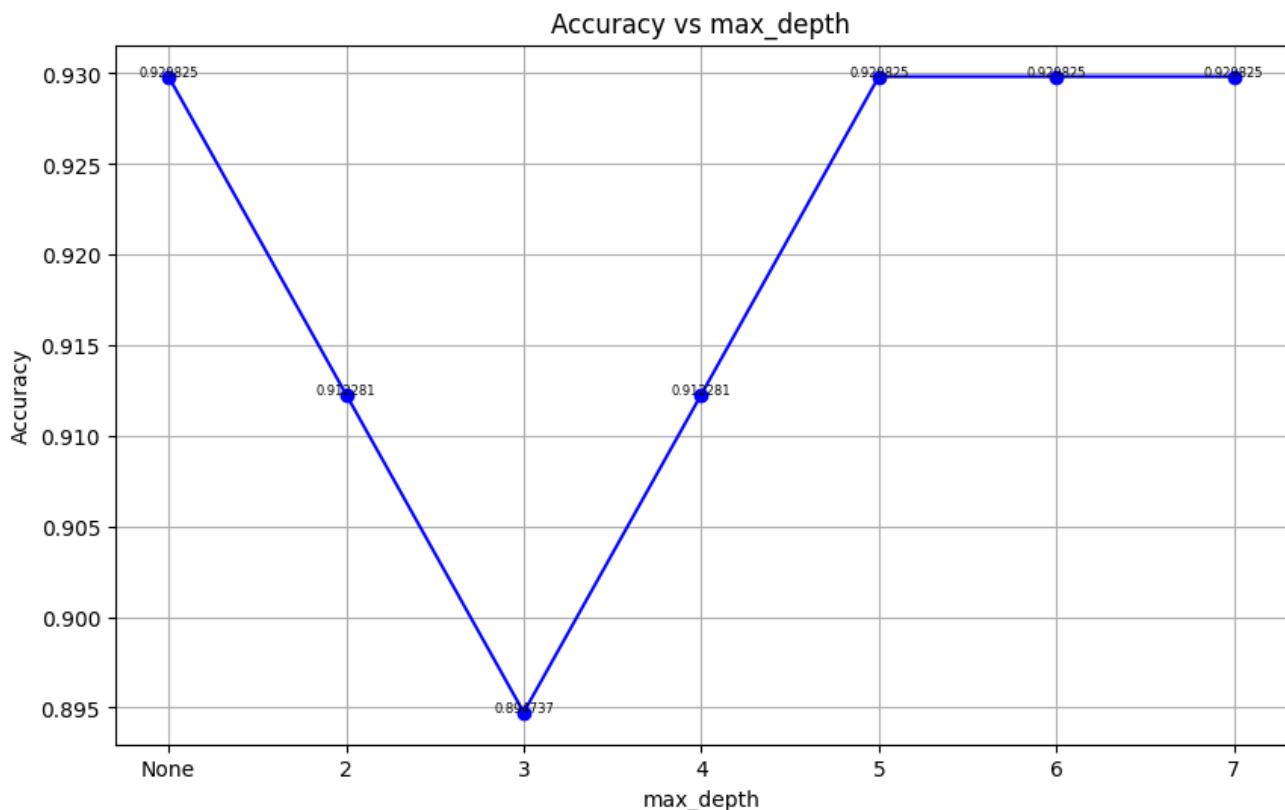


Ảnh cây quyết định với max_depth = 7



Bảng kết quả

max_depth	None	2	3	4	5	6	7
Accuracy	0.929825	0.912281	0.894737	0.912281	0.929825	0.929825	0.929825



- Độ Chính Xác Cao Ở Độ Sâu Tối Đa (max_depth = None): Khi không giới hạn độ sâu (max_depth=None), độ chính xác đạt giá trị cao nhất là 0.929825. Điều này cho thấy mô hình có thể học tốt dữ liệu mà không gặp vấn đề lớn về quá khớp.
- Ảnh hưởng của Độ Sâu: Độ chính xác giảm xuống mức thấp nhất (0.894737) tại max_depth=3, cho thấy độ sâu nhỏ có thể khiến cây quyết định không đủ khả năng học dữ liệu phức tạp. Khi tăng độ sâu từ max_depth=4 đến max_depth=7, độ chính xác dần tăng lên và giữ ổn định ở 0.929825, chứng minh rằng mô hình đạt cân bằng giữa việc học dữ liệu và tránh quá khớp.

4. Additional dataset (Mushroom)

4.1. Tập dữ liệu

Nguồn: [Mushroom - UCI Machine Learning Repository](#)

Thông tin: - Số mẫu: 8124

- Số đặc trưng: 23 (các thuộc tính hình thái)
- Mục tiêu: Phân loại nấm độc hoặc ăn được

Tập dữ liệu bao gồm các mẫu tả được mô tả theo hình thái, màu sắc, mùi hương, bề mặt, và các yếu tố khác giúp nhận diện và phân loại nấm thuộc 23 loài nấm trong họ Agaricus và Lepiota.

4.2. Chuẩn bị dữ liệu

Dữ liệu được đọc từ tập tin agaricus-lepiota.data bằng thư viện pandas.

Tiền xử lý:

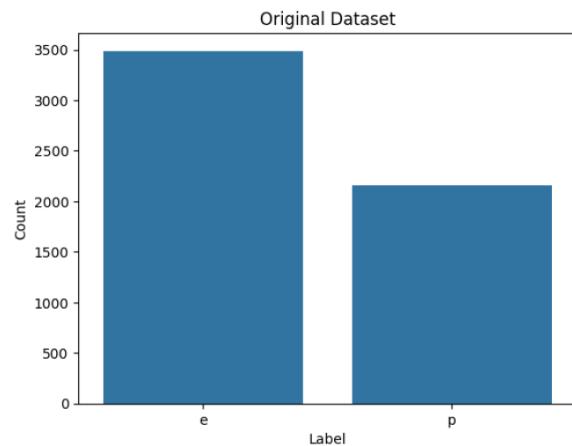
- Chuyển đổi dữ liệu danh mục sang dạng số bằng phương pháp one-hot encoding.
- Kiểm tra và xử lý giá trị khuyết (missing values) trong các đặc trưng (nếu có).
- Nhóm nhãn: Phân chia dữ liệu thành hai nhóm chính "ăn được" và "độc".
- Chuẩn hóa các đặc trưng bằng phương pháp giữ nguyên giá trị ban đầu (không cần chuẩn hóa do dữ liệu đã ở dạng danh mục).

Chia dữ liệu thành các tập:

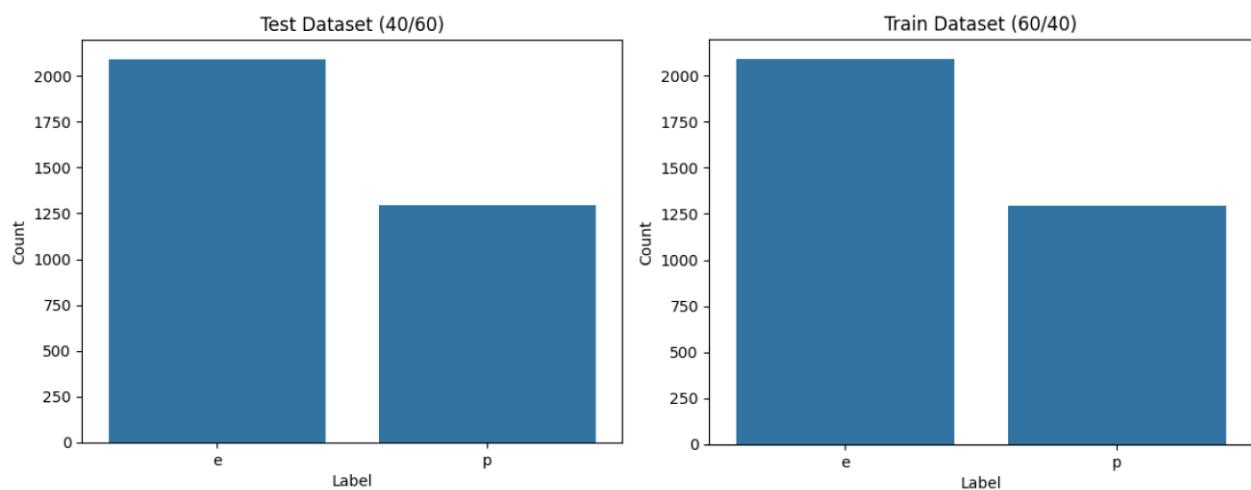
- Chia tập train/test theo các tỷ lệ 40/60, 60/40, 80/20, 90/10.
- Sử dụng phương pháp StratifiedShuffleSplit để giữ nguyên phân phối nhãn giữa các tập dữ liệu.

Trực quan hóa: Biểu đồ phân phối các nhãn (ăn được/độc) để quan sát sự cân bằng của dữ liệu.

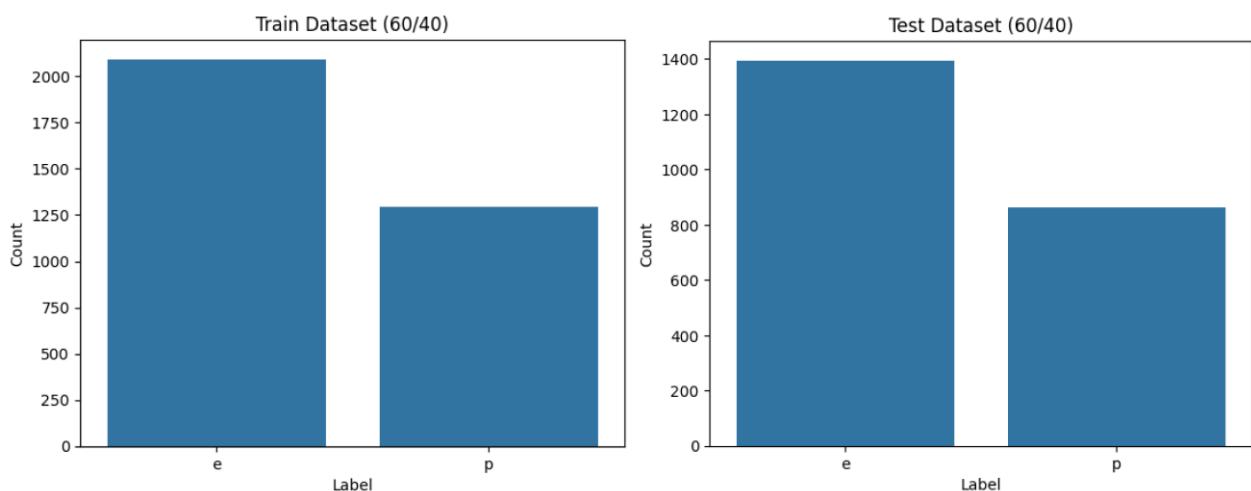
Dữ liệu gốc



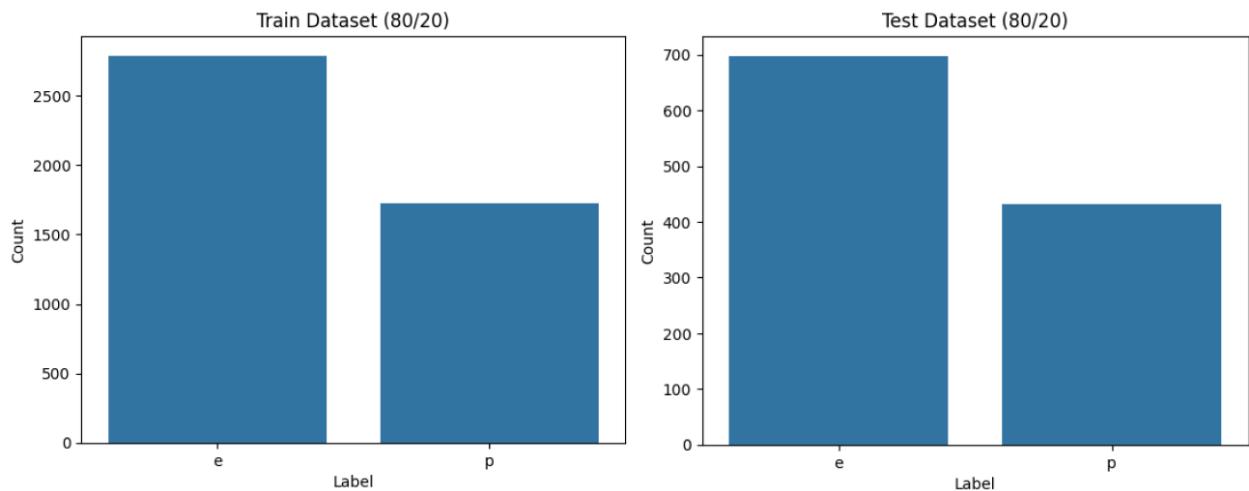
Tập train/test theo các tỷ lệ 40/60



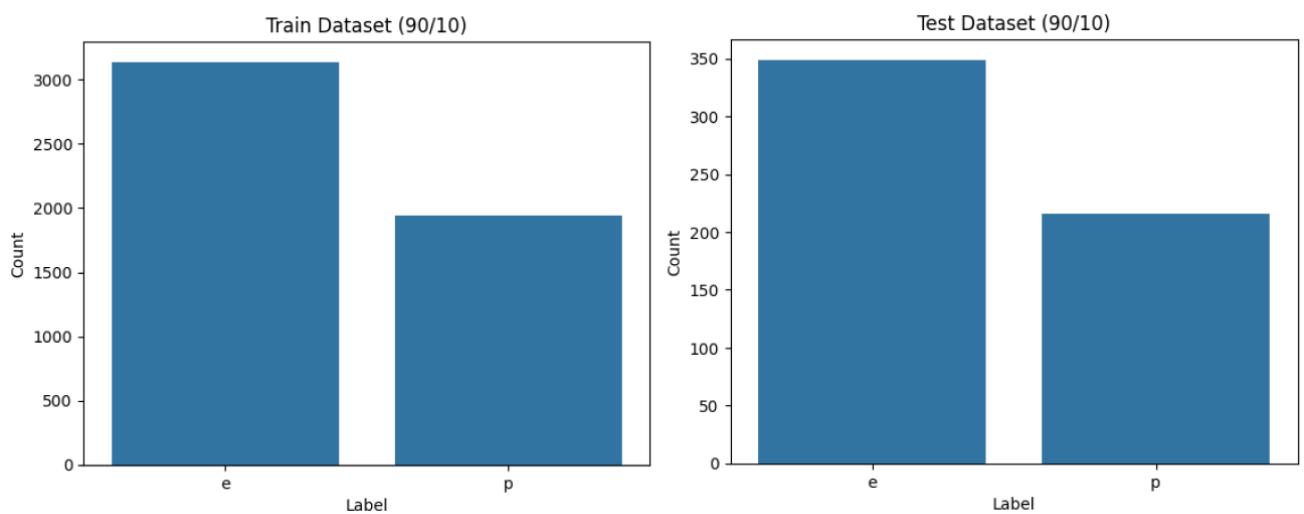
Tập train/test theo các tỷ lệ 60/40



Tập train/test theo các tỷ lệ 80/20



Tập train/test theo các tỷ lệ 90/10



4.3. Xây dựng mô hình cây quyết định

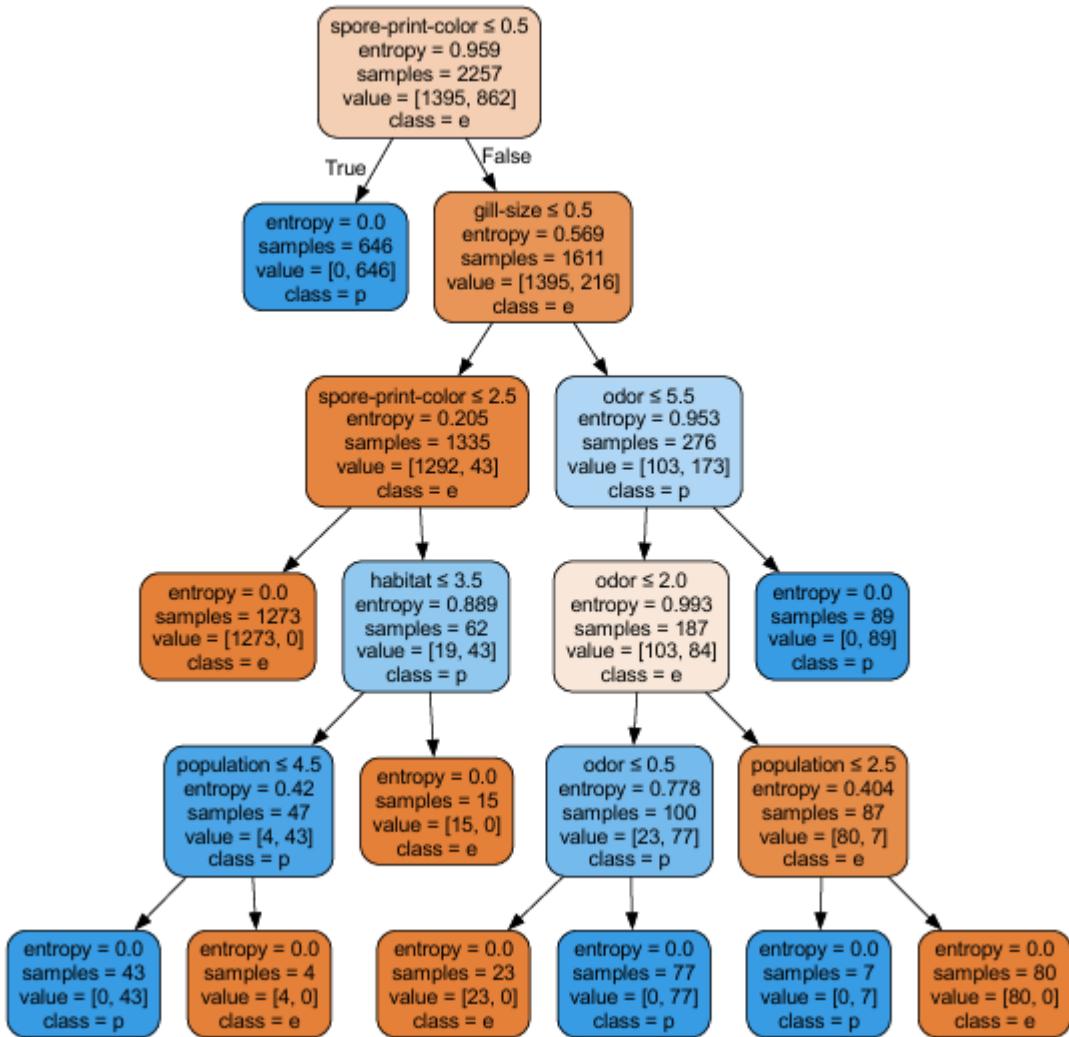
Công cụ sử dụng: DecisionTreeClassifier của thư viện scikit-learn.

Thông số quan trọng:

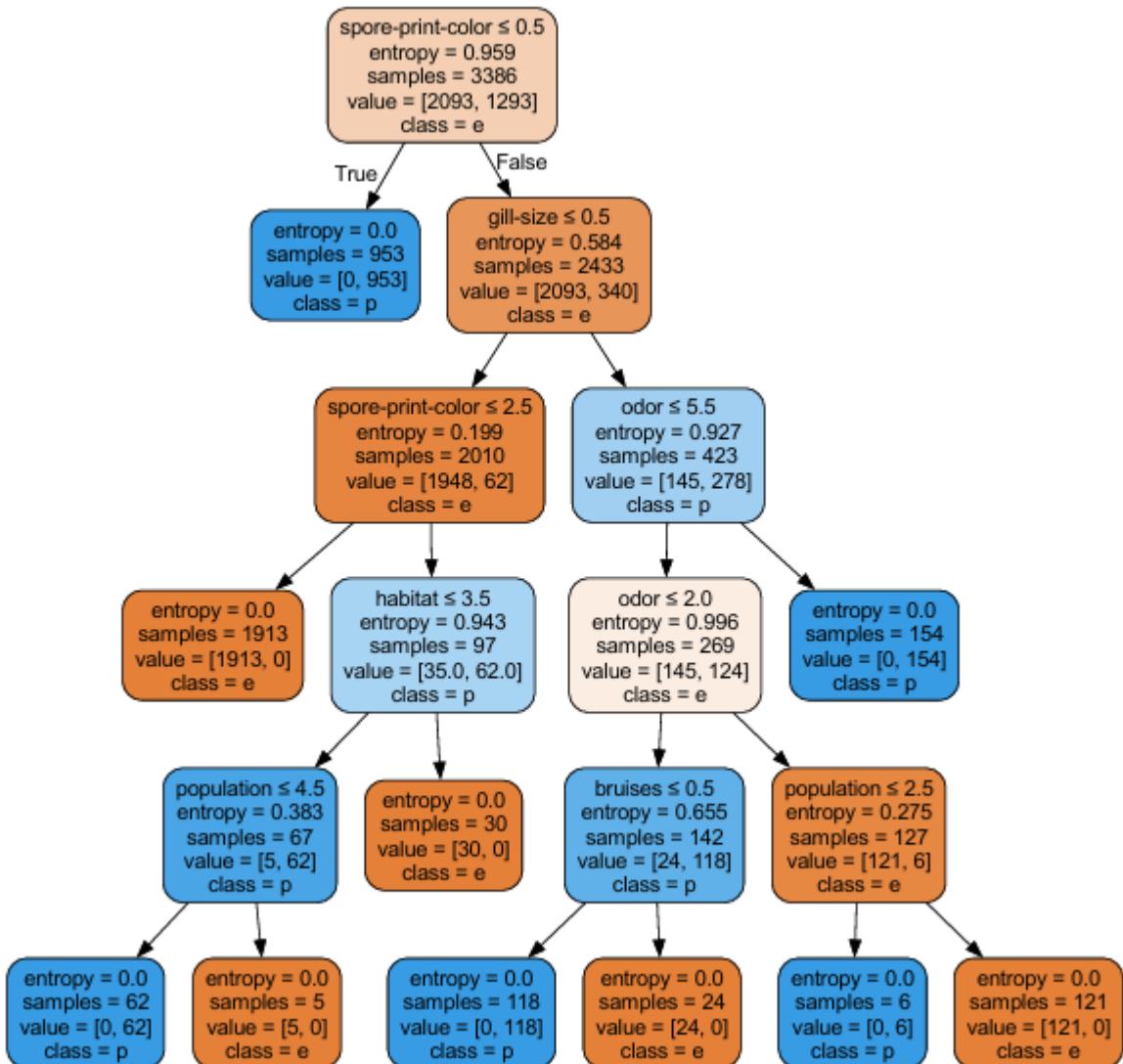
- Tiêu chí phân tách: Information Gain (sử dụng entropy).
- Độ sâu tối đa (max_depth): Được tùy chỉnh và thử nghiệm để tối ưu hóa mô hình.
- Sử dụng Graphviz để hiển thị cấu trúc cây.

Kết quả mong đợi: Cây quyết định được hiển thị tương ứng với từng tỷ lệ train/test.

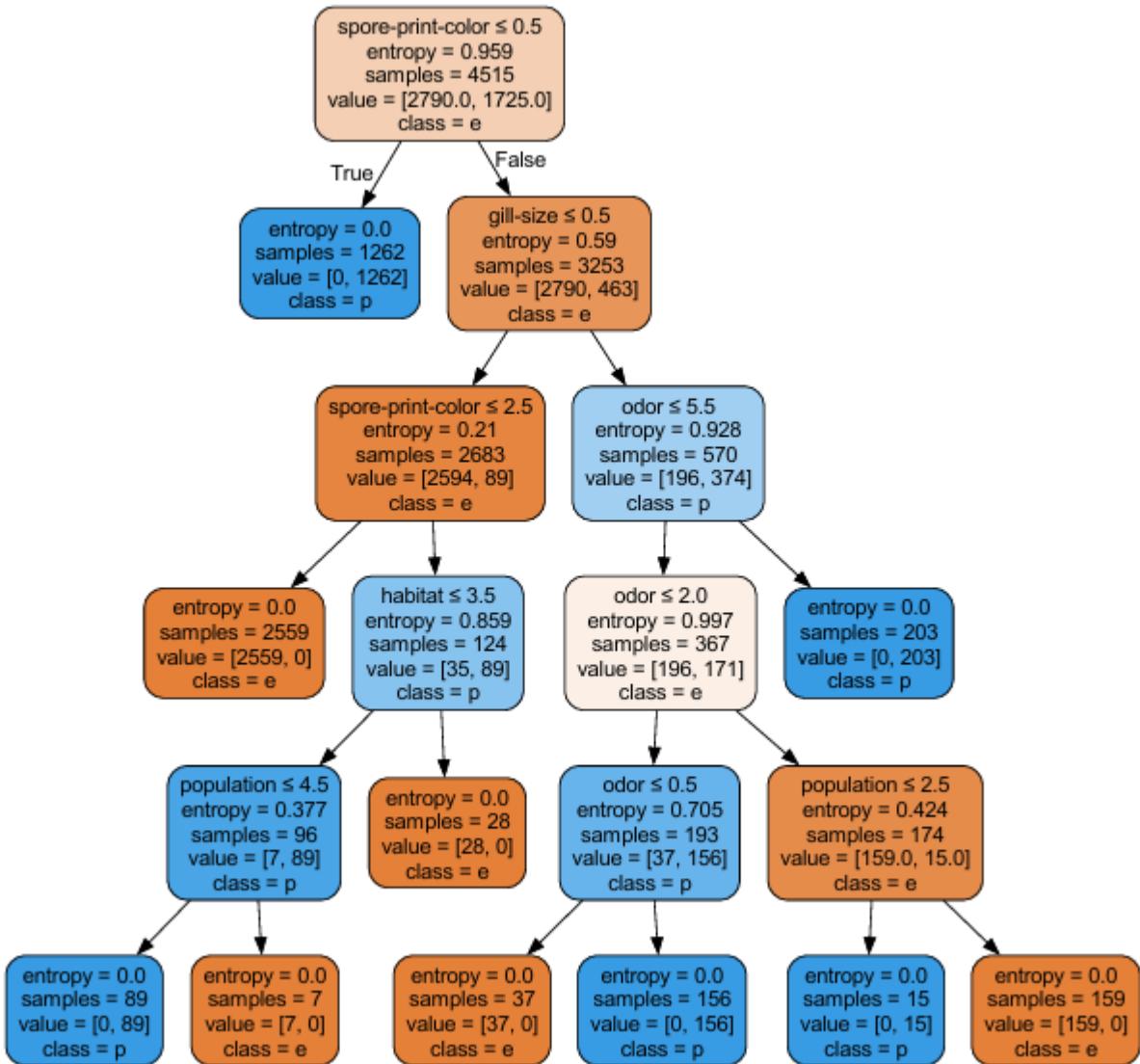
Tỉ lệ tập train/test là 40/60



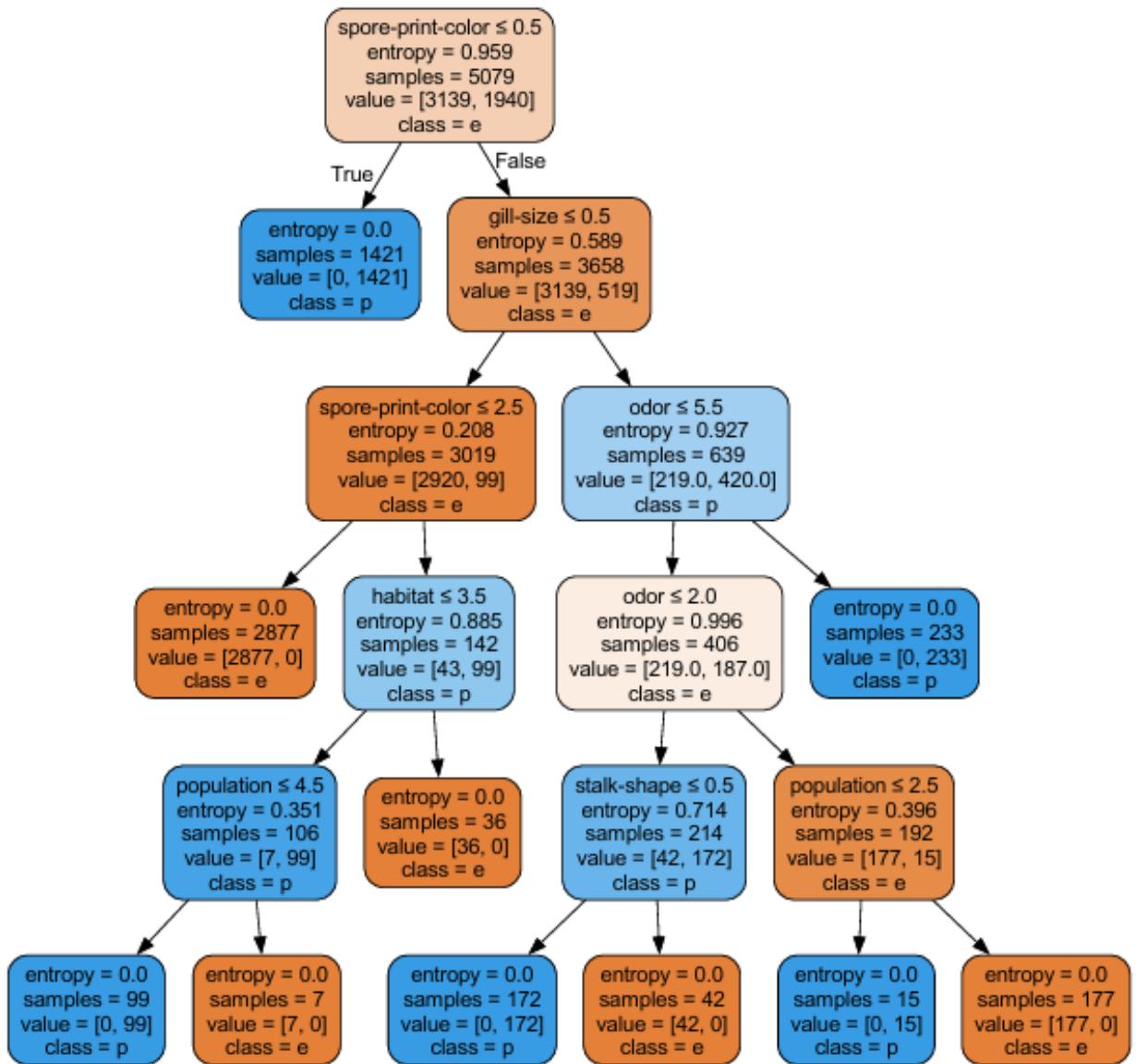
Tỉ lệ tập train/test là 60/40



Tỉ lệ tập train/test là 80/20



Tỉ lệ tập train/test là 90/10



4.4. Đánh giá mô hình cây quyết định

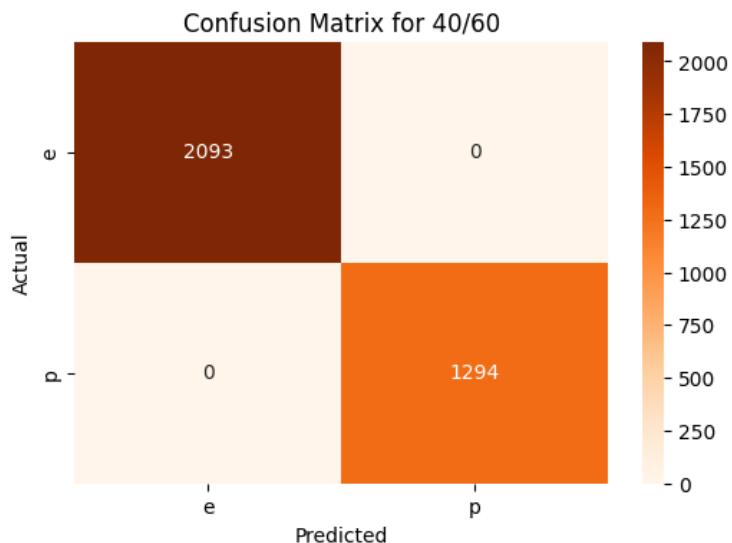
e: Edible (ăn được) - p: Poisonous (không ăn được)

Tỉ lệ tập train/test là 40/60

	precision	recall	f1-score	support
e	1.00	1.00	1.00	2093
p	1.00	1.00	1.00	1294
accuracy			1.00	3387
macro avg	1.00	1.00	1.00	3387
weighted avg	1.00	1.00	1.00	3387

Classification Report

Confusion Matrix



- Tỉ lệ chính xác: 100%
- 100% các mẫu mà mô hình dự đoán có nhãn 'e' là thực sự đúng với thực tế
- 100% các mẫu mà mô hình dự đoán có nhãn 'p' là thực sự đúng với thực tế
- Mô hình đã phát hiện được 100% mẫu có nhãn 'e'
- Mô hình đã phát hiện được 100% mẫu có nhãn 'p'
- Đối với nhãn 'e', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00
- Đối với nhãn 'p', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00

Đánh giá hiệu suất

Ở cả nhãn 'e' và 'p' mô hình đặt precision, recall và F1-score hoàn hảo là 1.00, cho thấy tất cả các mẫu 'e' và 'p' đều được dự đoán chính xác mà không có bất kỳ lỗi

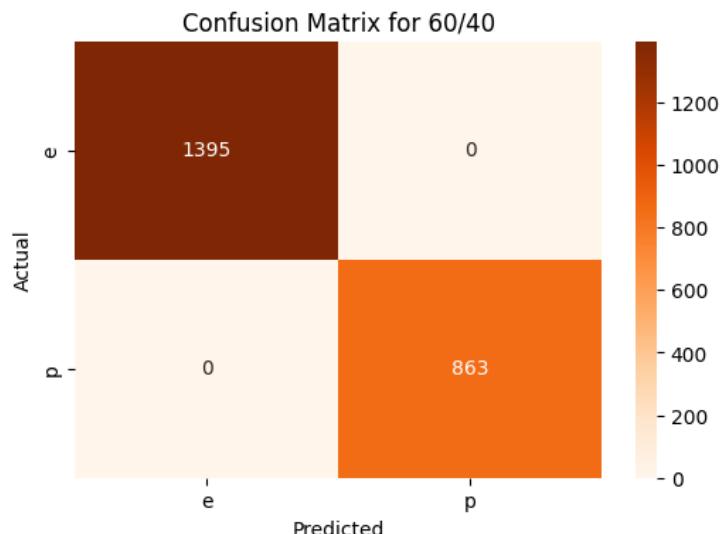
=> Điều này cho thấy mô hình xử lý rất tốt trong việc phân biệt các mẫu nấm độc với nấm ăn được.

Tỉ lệ tập train/test là 60/40

	precision	recall	f1-score	support
e	1.00	1.00	1.00	1395
p	1.00	1.00	1.00	863
accuracy			1.00	2258
macro avg	1.00	1.00	1.00	2258
weighted avg	1.00	1.00	1.00	2258

Classification Report

Confusion Matrix



- Tỉ lệ chính xác: 100%
- 100% các mẫu mà mô hình dự đoán có nhãn 'e' là thực sự đúng với thực tế
- 100% các mẫu mà mô hình dự đoán có nhãn 'p' là thực sự đúng với thực tế
- Mô hình đã phát hiện được 100% mẫu có nhãn 'e'
- Mô hình đã phát hiện được 100% mẫu có nhãn 'p'
- Đối với nhãn 'e', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00
- Đối với nhãn 'p', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00

Dánh giá hiệu suất

Ở cả nhãn 'e' và 'p' mô hình đặt precision, recall và F1-score hoàn hảo là 1.00, cho thấy tất cả các mẫu 'e' và 'p' đều được dự đoán chính xác mà không có bất kỳ lỗi

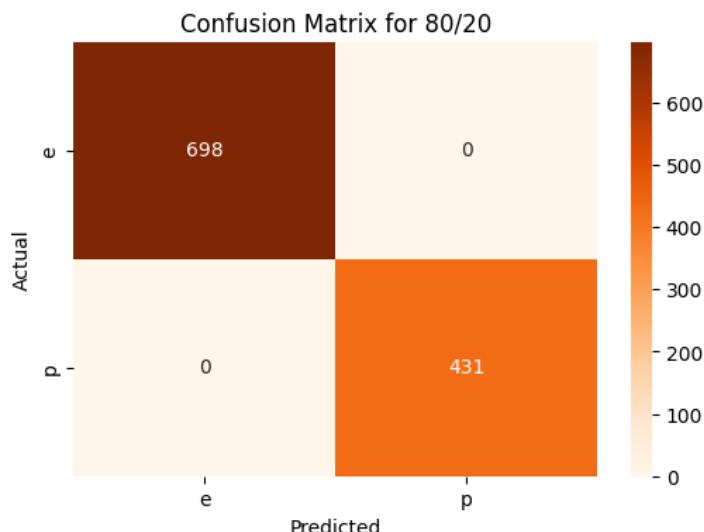
=> Điều này cho thấy mô hình xử lý rất tốt trong việc phân biệt các mẫu nấm độc với nấm ăn được.

Tỉ lệ tập train/test là 80/20

	precision	recall	f1-score	support
e	1.00	1.00	1.00	698
p	1.00	1.00	1.00	431
accuracy			1.00	1129
macro avg	1.00	1.00	1.00	1129
weighted avg	1.00	1.00	1.00	1129

Classification Report

Confusion Matrix



- Tỉ lệ chính xác: 100%
- 100% các mẫu mà mô hình dự đoán có nhãn 'e' là thực sự đúng với thực tế
- 100% các mẫu mà mô hình dự đoán có nhãn 'p' là thực sự đúng với thực tế
- Mô hình đã phát hiện được 100% mẫu có nhãn 'e'
- Mô hình đã phát hiện được 100% mẫu có nhãn 'p'
- Đối với nhãn 'e', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00
- Đối với nhãn 'p', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00

Dánh giá hiệu suất

Ở cả nhãn 'e' và 'p' mô hình đặt precision, recall và F1-score hoàn hảo là 1.00, cho thấy tất cả các mẫu 'e' và 'p' đều được dự đoán chính xác mà không có bất kỳ lỗi

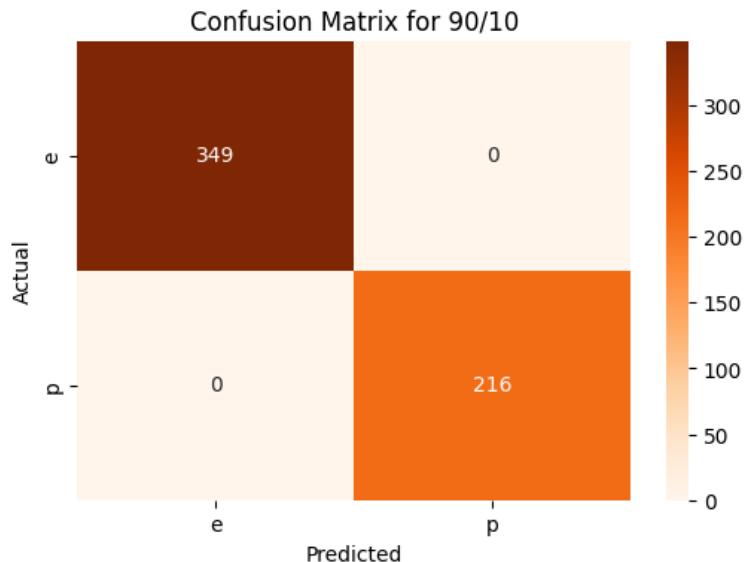
=> Điều này cho thấy mô hình xử lý rất tốt trong việc phân biệt các mẫu nấm độc với nấm ăn được.

Tỉ lệ tập train/test là 90/10

	precision	recall	f1-score	support
e	1.00	1.00	1.00	349
p	1.00	1.00	1.00	216
accuracy			1.00	565
macro avg	1.00	1.00	1.00	565
weighted avg	1.00	1.00	1.00	565

Classification Report

Confusion Matrix



- Tỉ lệ chính xác: 100%
- 100% các mẫu mà mô hình dự đoán có nhãn 'e' là thực sự đúng với thực tế
- 100% các mẫu mà mô hình dự đoán có nhãn 'p' là thực sự đúng với thực tế
- Mô hình đã phát hiện được 100% mẫu có nhãn 'e'
- Mô hình đã phát hiện được 100% mẫu có nhãn 'p'
- Đối với nhãn 'e', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00
- Đối với nhãn 'p', mô hình có chỉ số cân bằng giữa 2 chỉ số precision và recall là 1.00

Dánh giá hiệu suất

Ở cả nhãn 'e' và 'p' mô hình đặt precision, recall và F1-score hoàn hảo là 1.00, cho thấy tất cả các mẫu 'e' và 'p' đều được dự đoán chính xác mà không có bất kỳ lỗi

=> Điều này cho thấy mô hình xử lý rất tốt trong việc phân biệt các mẫu nấm độc với nấm ăn được.

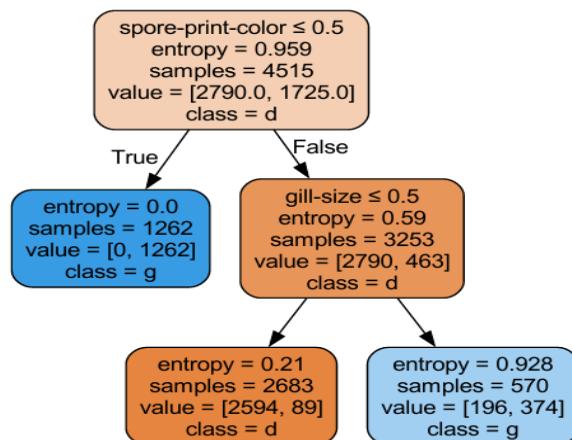
4.5. Ảnh hưởng của độ sâu đối với độ chính xác của cây quyết định

- Biểu đồ mô hình cây ứng với từng max depth

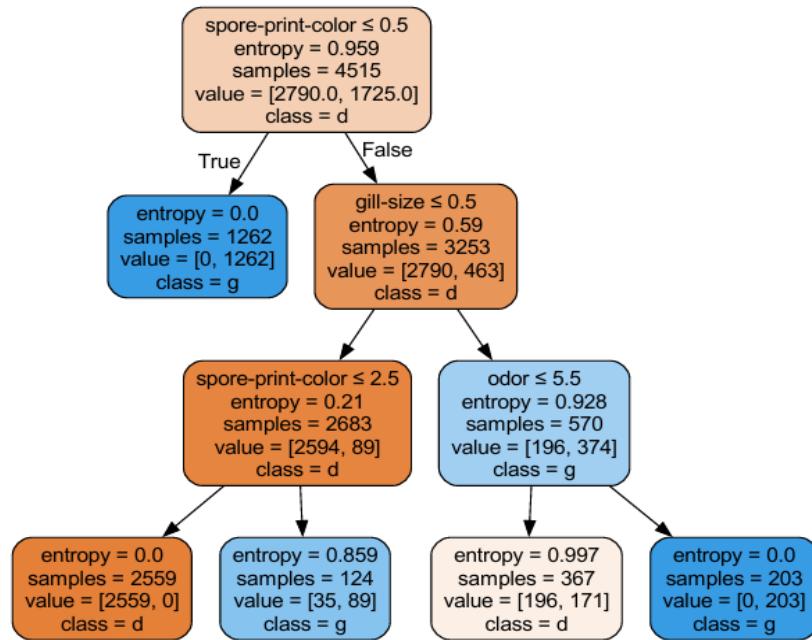
Đối với max depth = None



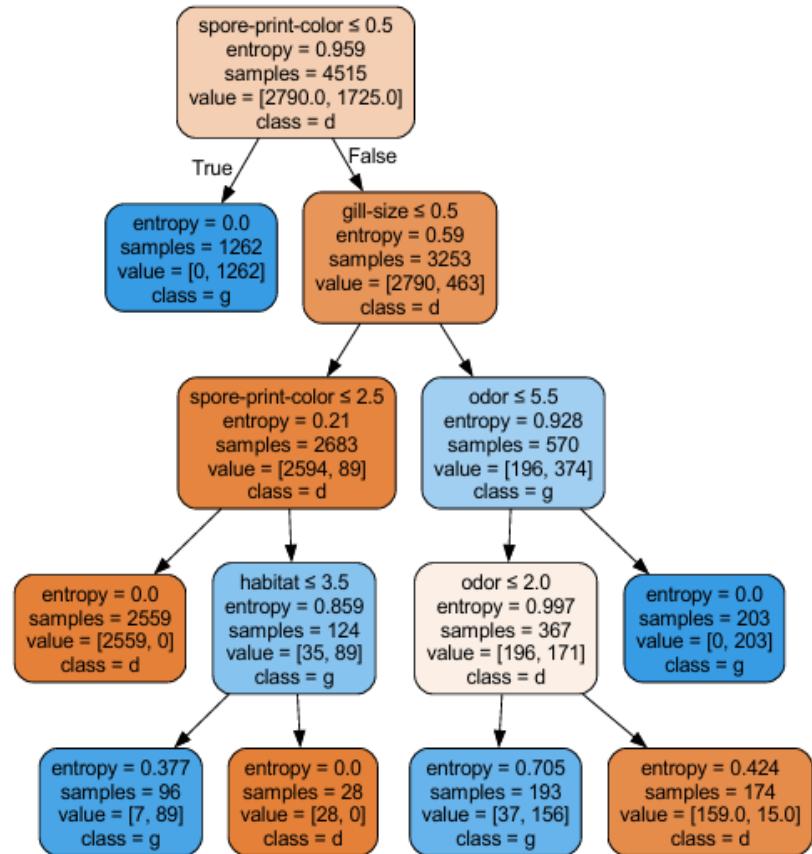
Đối với max depth = 2



Đối với max depth = 3



Đối với max depth = 4



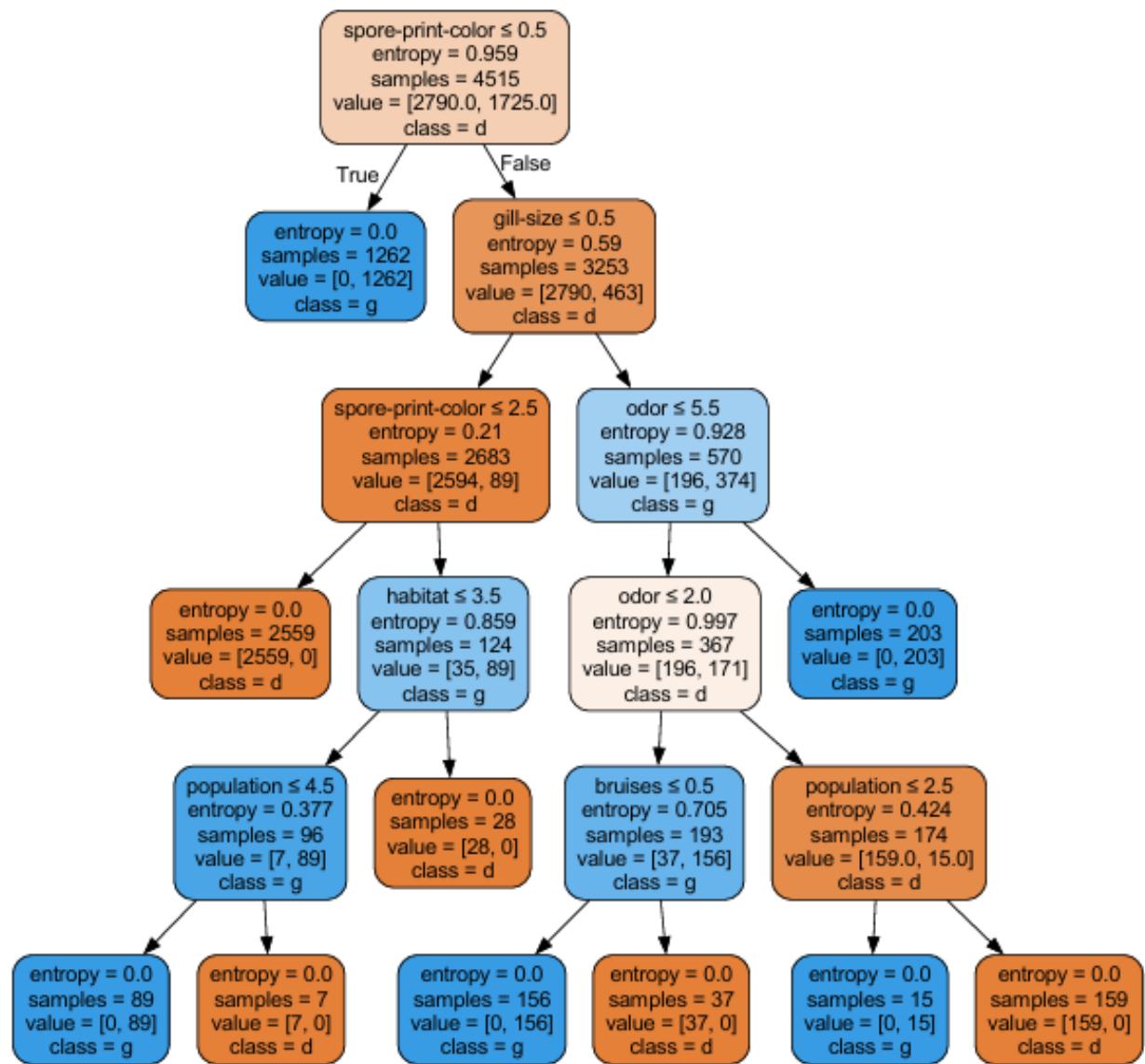
Đối với max depth = 5



Đối với max depth = 6



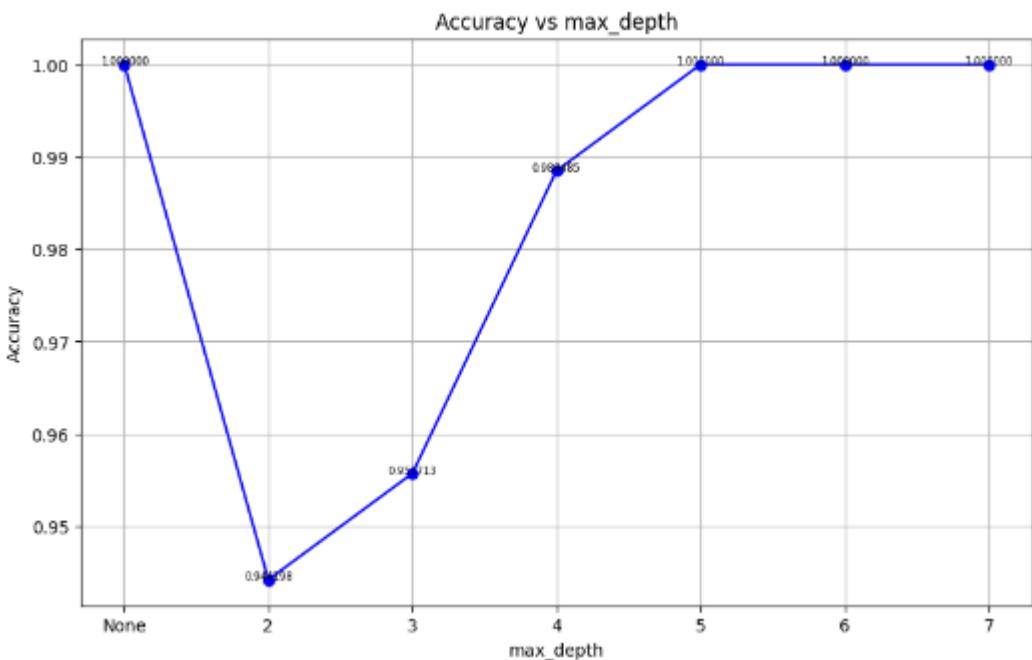
Đối với max depth = 7



Bảng kết quả:

max_depth	None	2	3	4	5	6	7
Accuracy	1.000000	0.944198	0.955713	0.988485	1.000000	1.000000	1.000000

Biểu đồ độ chính xác



- Khi không giới hạn độ sâu của cây ($\text{max_depth} = \text{None}$), mô hình đạt độ chính xác (accuracy) cao nhất là 1.00. Điều này cho thấy cây có thể học tất cả các mối quan hệ trong dữ liệu.
- Khi giới hạn độ sâu của cây ở giá trị nhỏ như 2 hoặc 3, độ chính xác giảm mạnh (xuống khoảng 0.94 và 0.95). Đây là hiện tượng underfitting, do cây không đủ phức tạp để học đầy đủ mối quan hệ trong dữ liệu.
- Ở độ sâu là 4, độ chính xác tăng đáng kể và gần đạt giá trị tối đa (0.99).
- Ở $\text{max_depth} = 5$, mô hình đạt lại độ chính xác hoàn hảo (1.00), cho thấy cây đã đủ phức tạp để học toàn bộ thông tin từ dữ liệu mà không bị underfitting.
- Ở độ sâu 6 và 7, độ chính xác vẫn duy trì ở mức 1.00. Điều này cho thấy việc tăng thêm độ sâu không cải thiện độ chính xác, đồng thời có nguy cơ dẫn đến overfitting nếu dữ liệu bị nhiễu hoặc không cân đối.

5. So sánh trên 3 tập dữ liệu

5.1 Đặc điểm dữ liệu

Chỉ tiêu	Mushroom	Breast Cancer Wisconsin	Wine Quality
Loại bài toán	Phân loại nhị phân	Phân loại nhị phân	Phân loại đa lớp
Số lượng mẫu	8124	569	4898
Số lượng lớp (Labels)	2 (Edible,Poisonous)	2 (Benign, Malignant)	3 (Low, Medium, High)
Tỉ lệ phân phối nhãn (Labels)	Cân bằng: - Edible = 4208, - Poisonous = 3916	Tương đối cân bằng: - Benign = 357 (63%) - Malignant = 212 (37%)	Không cân bằng: - Medium (~70%) - Low (~15%) - High (~15%)
Số lượng đặc trưng (Features)	22 thuộc tính dạng categorical	30 thuộc tính dạng numerical	11 thuộc tính dạng numerical
Thuộc tính nổi bật	odor, gill-size,	mean radius, mean texture, worst area	alcohol, density, pH
Độ phức tạp của mối quan hệ	Rõ ràng, có thuộc tính quyết định mạnh	Tương đối phức tạp, mối quan hệ số học rõ ràng	Khá phức tạp, nhiều thuộc tính không tuyến tính.

5.2 Độ phức tạp của mô hình

Chỉ tiêu	Mushroom	Breast Cancer Wisconsin	Wine Quality
Độ sâu tối ưu	3-5	4-6	6-7
Overfitting	Ít xảy ra	Trung bình	Dễ xảy ra ở depth cao
Tốc độ huấn luyện	Nhanh	Trung bình	Chậm
Độ phức tạp nhánh	Đơn giản	Phức tạp trung bình	Phức tạp cao
Phân tích	Với độ sâu nhỏ (3-5), cây quyết định đạt hiệu quả tối ưu do Cân độ sâu trung bình (4-6) để phân tích chính xác dữ liệu số phức tạp. Dữ liệu phân lớp rõ ràng.	Cần độ sâu trung bình (4-6) để phân tích chính xác dữ liệu số phức tạp	Cần độ sâu lớn hơn (6-7), nhưng dễ gặp overfitting do dữ liệu không cân bằng và đa lớp.

5.3 Hiệu năng thực tế

Chỉ tiêu	Mushroom	Breast Cancer Wisconsin	Wine Quality
Độ chính xác (Accuracy)	98%	96-97%	65-75%
F1-Score (Macro)	1.00	0.97	0.70
Thời gian suy diễn	Nhanh	Trung bình	Chậm

5.4 So sánh tổng quan

Chỉ tiêu	Mushroom	Breast Cancer Wisconsin	Wine Quality
Độ phức tạp dữ liệu	Thấp	Trung bình	Cao
Khả năng phân tách tự nhiên	Cao	Trung bình	Thấp
Hiệu năng cây quyết định	Xuất sắc (~100%)	Rất tốt (~97-98%)	Khá (65-75%)
Ứng dụng thực tiễn	Tốt (phát hiện nấm độc)	Tốt (phân tích ung thư)	Thách thức (phân loại rượu)
Kết luận	<p>Phù hợp cho minh họa cây quyết định, đặc biệt khi muốn giải thích trực quan</p> <p>Dễ đạt hiệu năng cao nhất (100%) do dữ liệu rõ ràng và cân bằng.</p>	<p>Dữ liệu thực tế, phù hợp để kiểm tra mô hình trong các bài toán nhị phân phức tạp hơn.</p> <p>Yêu cầu tinh chỉnh độ sâu để đạt hiệu năng tối ưu (~97-98%).</p>	<p>Là bài toán đa lớp điển hình, có nhiều thách thức hơn, đòi hỏi tối ưu hóa tốt.</p> <p>Hiệu năng thấp hơn, nhưng cung cấp cơ hội để đánh giá khả năng tổng quát hóa của cây quyết định.</p>

Tài liệu tham khảo

- [1] Turgay, G. (2020) *Comparing classification models for Wine Quality Prediction*, Medium. Available at: <https://towardsdatascience.com/comparing-classification-models-for-wine-quality-prediction-6c5f26669a4f> .
- [2] Hackers Realm (2023) *Wine quality prediction analysis using Python: Classification: Machine learning project tutorial*, Hackers Realm. Available at: https://www.hackersrealm.net/post/wine-quality-prediction-analysis-using-python#google_vignette .
- [3] Aras, M. (2023) *Breast cancer classification with decision tree*, Medium. Available at: <https://medium.com/@aras.merve.20/breast-cancer-classification-with-decision-tree-f137052fdd39> .
- [3] Tank, K. (2020) *Mushroom classification using different classifiers*, Medium. Available at: <https://medium.com/analytics-vidhya/mushroom-classification-using-different-classifiers-aa338c1cd0ff> .