

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

----- -----



**BÁO CÁO
ĐỒ ÁN CƠ SỞ TRÍ TUỆ NHÂN TẠO
CÂY QUYẾT ĐỊNH**

Giáo viên hướng dẫn: Nguyễn Thanh Tịnh
Sinh viên thực hiện: 22120117- Trần Mạnh Hùng
22120339 – Nguyễn Thị Anh Thi
22120416 – Huỳnh Thị Kim Tuyền
22120421 – Nguyễn Đoàn Minh Uyên

Hồ Chí Minh, tháng 12 năm 2024

MỤC LỤC

I. Tổng quan đồ án	4
1. Giới thiệu về đồ án	4
2. Mục tiêu của đồ án.....	4
II. Phân công công việc	5
Chi tiết phân công công việc.....	5
III. Đánh giá mức độ hoàn thành đồ án.....	5
IV. CÁC THU VIỆN SỬ DỤNG.....	5
V. Nội dung đồ án.....	6
1. Giới thiệu các dataset.....	6
2. Phân tích các dataset	7
2.1 Breast Cancer Dataset.....	7
2.1.1. Chuẩn bị dữ liệu	7
2.1.2 Cài đặt.....	8
2.1.3 Trực quan hóa cây quyết định	9
2.1.4 Đánh giá cây quyết định	12
2.1.5 Phân tích độ sâu và độ chính xác	21
2.2 Wine Quality Dataset.....	26
2.2.1 Chuẩn bị dữ liệu	26
2.2.2 Cài đặt.....	30
2.2.3 Trực quan hóa cây quyết định	30
2.2.4 Đánh giá cây quyết định	31
2.2.5 Phân tích độ sâu và độ chính xác	41
2.3 Additional Dataset - Car Purchase Dataset	44
2.3.1 Chuẩn bị dữ liệu.....	44
2.3.2 Cài đặt.....	45
2.3.3 Trực quan hóa cây quyết định	46
2.3.4 Đánh giá cây quyết định	47
2.3.5 Phân tích độ sâu và độ chính xác	56
3. So sánh 3 dataset.....	59
4. Phân tích các đặc điểm ảnh hưởng đến cây quyết định của các dataset..	59
4.1 Số lượng lớp	59
4.2 Số đặc trưng.....	60

4.3 Kích thước mẫu	60
4.4 Kết luận.....	60
TÀI LIỆU THAM KHẢO.....	61

THÔNG TIN THÀNH VIÊN NHÓM

STT	MSSV	HỌ VÀ TÊN
1	22120117	Trần Mạnh Hùng
2	22120339	Nguyễn Thị Anh Thi
3	22120416	Huỳnh Thị Kim Tuyền
4	22120421	Nguyễn Đoàn Minh Uyên

I. TỔNG QUAN ĐỒ ÁN

1. Giới thiệu về đồ án

- Trong đồ án này, nhóm sẽ triển khai và xây dựng mô hình cây quyết định (Decision Tree) để giải quyết các bài toán phân loại sử dụng thư viện scikit-learn trong Python.
- Cây quyết định là một thuật toán học máy phổ biến được sử dụng trong nhiều bài toán phân loại, đặc biệt là trong các trường hợp dữ liệu có cấu trúc rõ ràng và dễ hiểu. Đối với đồ án này, em sẽ áp dụng cây quyết định lên ba bộ dữ liệu thực tế: UCI Breast Cancer Wisconsin, UCI Wine Quality, Bộ dữ liệu bổ sung là Car Purchase.

2. Mục tiêu của đồ án

- Mục tiêu của đồ án là áp dụng và đánh giá hiệu quả của cây quyết định trong các bài toán phân loại. Nhóm sẽ biết cách xây dựng mô hình, đánh giá chất lượng mô hình thông qua các chỉ số như độ chính xác, và so sánh kết quả giữa các bộ dữ liệu. Bằng cách này, nhóm có thể hiểu sâu hơn về cách thức hoạt động của thuật toán cây quyết định và áp dụng nó vào các bài toán thực tế.

II. PHÂN CÔNG CÔNG VIỆC

Chi tiết phân công công việc

STT	Họ tên	Phân công	Tiến độ
1	Trần Mạnh Hùng	Additional dataset	100%
2	Nguyễn Thị Anh Thi	Wine Quality dataset	100%
3	Huỳnh Thị Kim Tuyền	Báo cáo + Review đồ án	100%
4	Nguyễn Đoàn Minh Uyên	Breast Cancer dataset.	100%

III. ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH ĐỒ ÁN

- Hoàn thành đủ các yêu cầu được đưa ra cho các tập dữ liệu bao gồm chuẩn bị data, cài đặt cây phân lớp, đánh giá hiệu suất của cây, phân tích độ sâu và mức độ chính xác tương ứng.
- Có phân tích và so sánh giữa 3 tập dữ liệu
- Tổ chức code hợp lý, có diễn giải để mô tả thêm.

IV. CÁC THU VIỆN SỬ DỤNG

1. scikit-learn: Thư viện chính để xây dựng và huấn luyện mô hình cây quyết định.
 - Cài đặt: pip install scikit-learn
2. numpy: Xử lý các mảng và phép toán đại số tuyến tính.
 - Cài đặt: pip install numpy
3. pandas: Quản lý và xử lý dữ liệu bảng.
 - Cài đặt: pip install pandas
4. matplotlib: Thư viện vẽ biểu đồ để trực quan hóa kết quả.
 - Cài đặt: pip install matplotlib
5. seaborn: Để dàng tạo các biểu đồ đẹp mắt, hỗ trợ Matplotlib.
 - Cài đặt: pip install seaborn
6. graphviz: Để trực quan hóa cây quyết định.
 - Cài đặt phần mềm: Tải và cài đặt Graphviz từ [graphviz.gitlab.io](https://graphviz.gitlab.io/download/).
 - Cài đặt thư viện Python: pip install graphviz
7. ucimlrepo : Để import dataset Breast Cancer
 - Cài đặt : pip install ucimlrepo

V. NỘI DUNG ĐỒ ÁN

1. Giới thiệu các dataset

- Breast Cancer Dataset : Gồm 569 mẫu với 30 đặc trưng và nhãn phân loại:
 - M (Malignant): Ác tính.
 - B (Benign): Lành tính.
- Wine Quality Dataset : Gồm 4898 mẫu với 11 đặc trưng, nhãn phân loại chia thành 10 mức chất lượng (0-10).
 - Gom nhóm nhãn thành:
 - Low Quality (0-4).
 - Standard Quality (5-6).
 - High Quality (7-10).
- Additional Dataset: Tập dữ liệu chứa thông tin về khách hàng và quyết định mua xe. Gồm 1000 mẫu với 3 đặc trưng gồm Gender, Age, AnnualSalary, nhãn phân loại là dạng nhị phân:
 - 0 : No Buy
 - 1 : Buy

2. Phân tích các dataset

2.1 Breast Cancer Dataset

2.1.1. Chuẩn bị dữ liệu

- Bộ dữ liệu sử dụng trong dự án này là Breast Cancer Wisconsin Diagnostic Dataset, được tải bằng hàm `fetch_ucirepo` từ module `ucimlrepo`. Bộ dữ liệu chứa thông tin liên quan đến đặc điểm tế bào trong các mẫu sinh thiết, nhằm phân loại khối u là lành tính (Benign) hoặc ác tính (Malignant).
- Thông tin chi tiết:
 - o Số đặc trưng (Features): 30.
 - o Số mẫu (Samples): 569.
 - o Đầu ra (Targets): Nhãn phân loại khối u:
 - B: Lành tính (Benign).
 - M: Ác tính (Malignant).

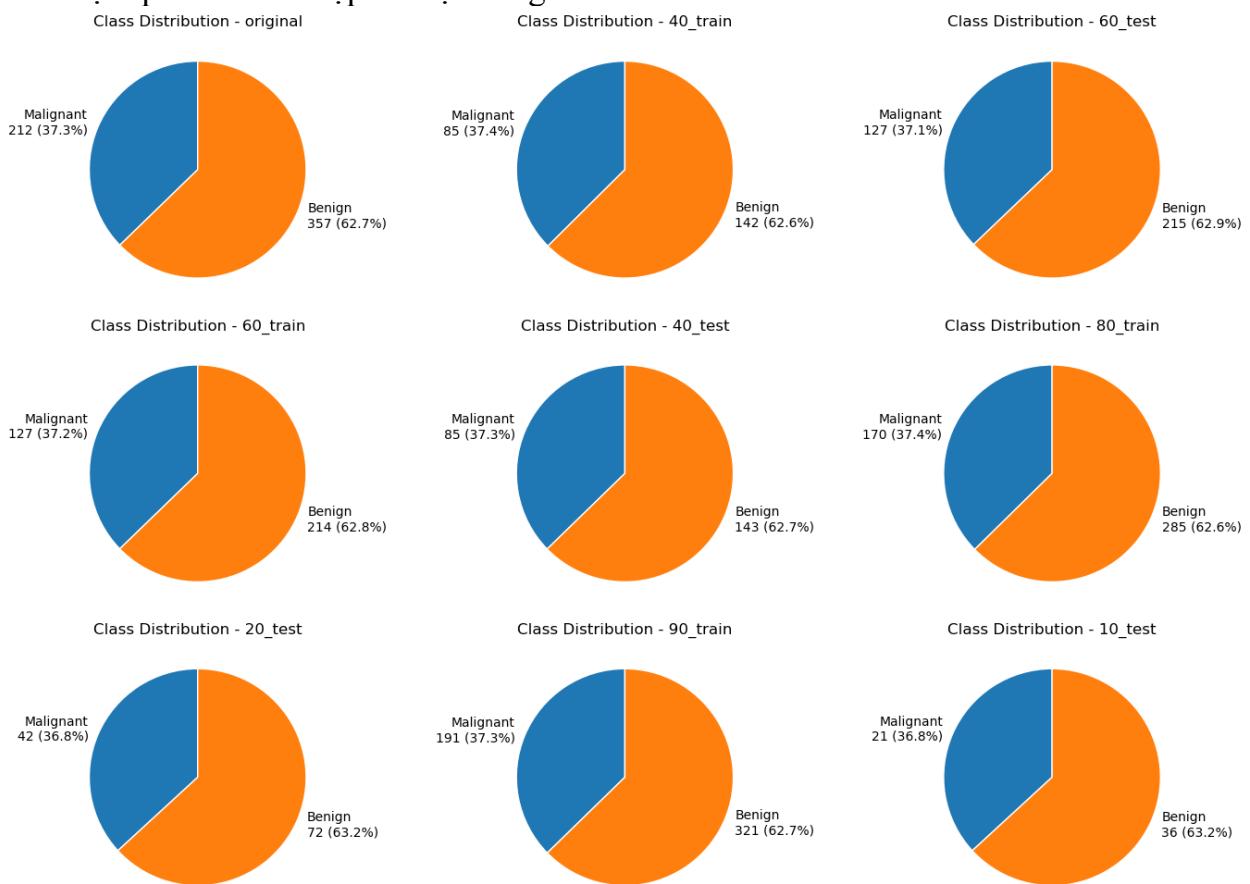
	radius3	texture3	perimeter3	area3	smoothness3	compactness3	concavity3	concave_points3	symmetry3	fractal_dimension3
...	25.38	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890
...	24.99	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
...	23.57	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
...	14.91	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
...	22.54	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

	radius1	texture1	perimeter1	area1	smoothness1	compactness1	concavity1	concave_points1	symmetry1	fractal_dimension1
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883

5 rows × 30 columns

- Bộ dữ liệu được tải vào cấu trúc DataFrame của thư viện pandas để dễ dàng thao tác. Tên các đặc trưng và tên các nhãn được trích xuất, đồng thời nhãn mục tiêu được chuyển đổi từ dạng chữ 'M' (Malignant - Ác tính) , 'B' (Benign - Lành tính) sang dạng số, tương ứng là 0 và 1, sử dụng hàm `map`.
- Dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo các tỷ lệ khác nhau (40%, 60%, 80%, và 90%) để đánh giá hiệu suất của mô hình dưới các kích thước tập huấn luyện khác nhau. Hàm `train_test_split` từ thư viện `sklearn.model_selection` được sử dụng cho mục đích này, với điều kiện chia `stratify` để đảm bảo phân phối lớp trong các tập dữ liệu được giữ nguyên.

- Trực quan hóa các tập dữ liệu bằng biểu đồ như sau



Nhận xét:

- Phân phối gốc (Original Dataset): Tập dữ liệu ban đầu có tỷ lệ là Malignant: 37.3% và Benign: 62.7%. Phân phối này không cân bằng, nhưng cũng không quá chênh lệch nghiêm trọng.
- Phân phối sau khi chia các tập con: Tất cả các tập con đều giữ nguyên tỷ lệ gần bằng với tập dữ liệu gốc. Điều này cho thấy việc chia dữ liệu đã giữ được tính đại diện của tập gốc, đảm bảo được mỗi tập con không bị lệch phân phối.

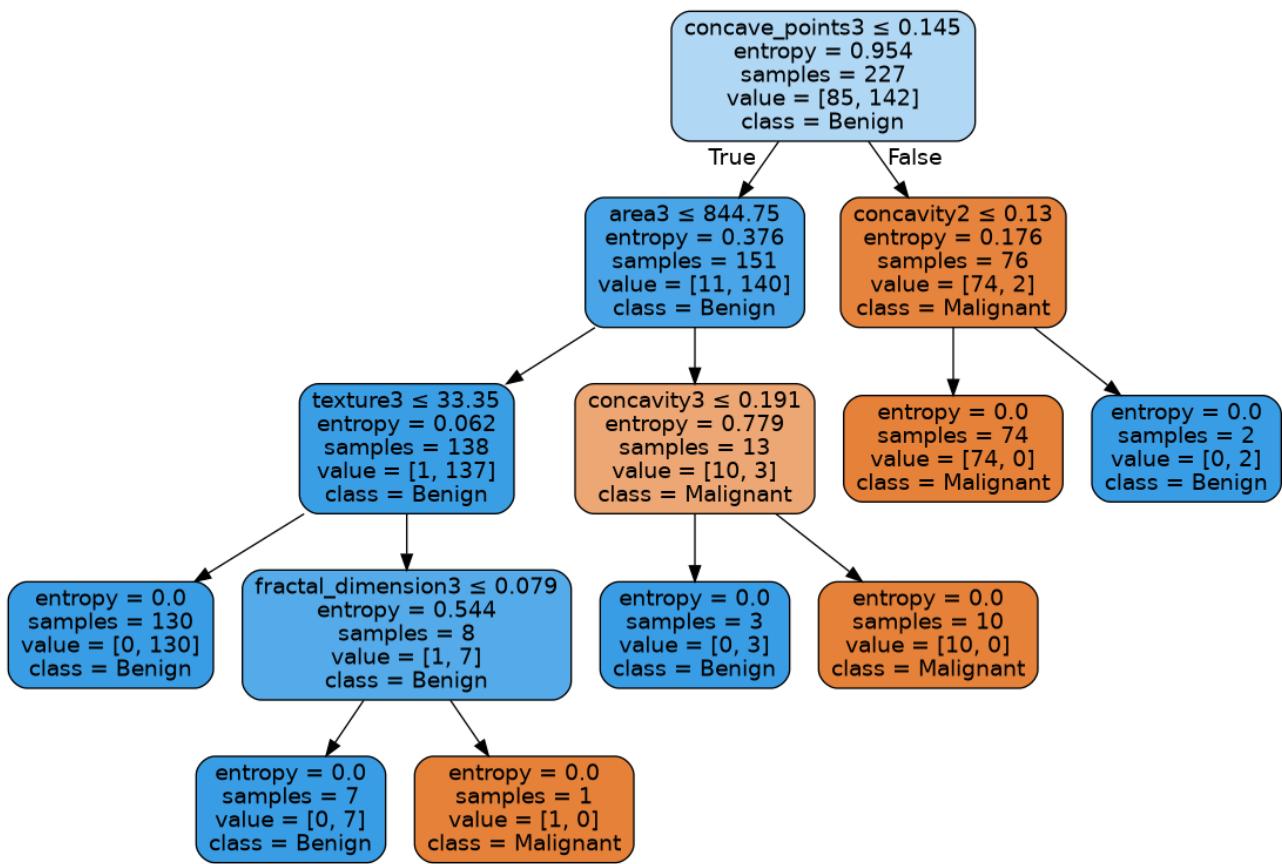
2.1.2 Cài đặt

- Sử dụng DecisionTreeClassifier từ scikit-learn để xây dựng và huấn luyện cây.
- Sử dụng export_graphviz: Xuất cấu trúc cây với thông tin đặc trưng, ngưỡng phân chia, và phân phối lớp.
- Sử dụng graphviz: Chuyển mã DOT thành hình ảnh cây với các nút làm tròn và tô màu theo tỷ lệ lớp.
- Nội dung hiển thị trên cây:
 - o Nút phân chia: đặc trưng, ngưỡng phân chia tại của đặc trưng, entropy, số lượng mẫu, phân phối lớp, nhãn.
 - o Nút lá: entropy, số lượng mẫu, phân phối lớp, nhãn.

2.1.3 Trực quan hóa cây quyết định

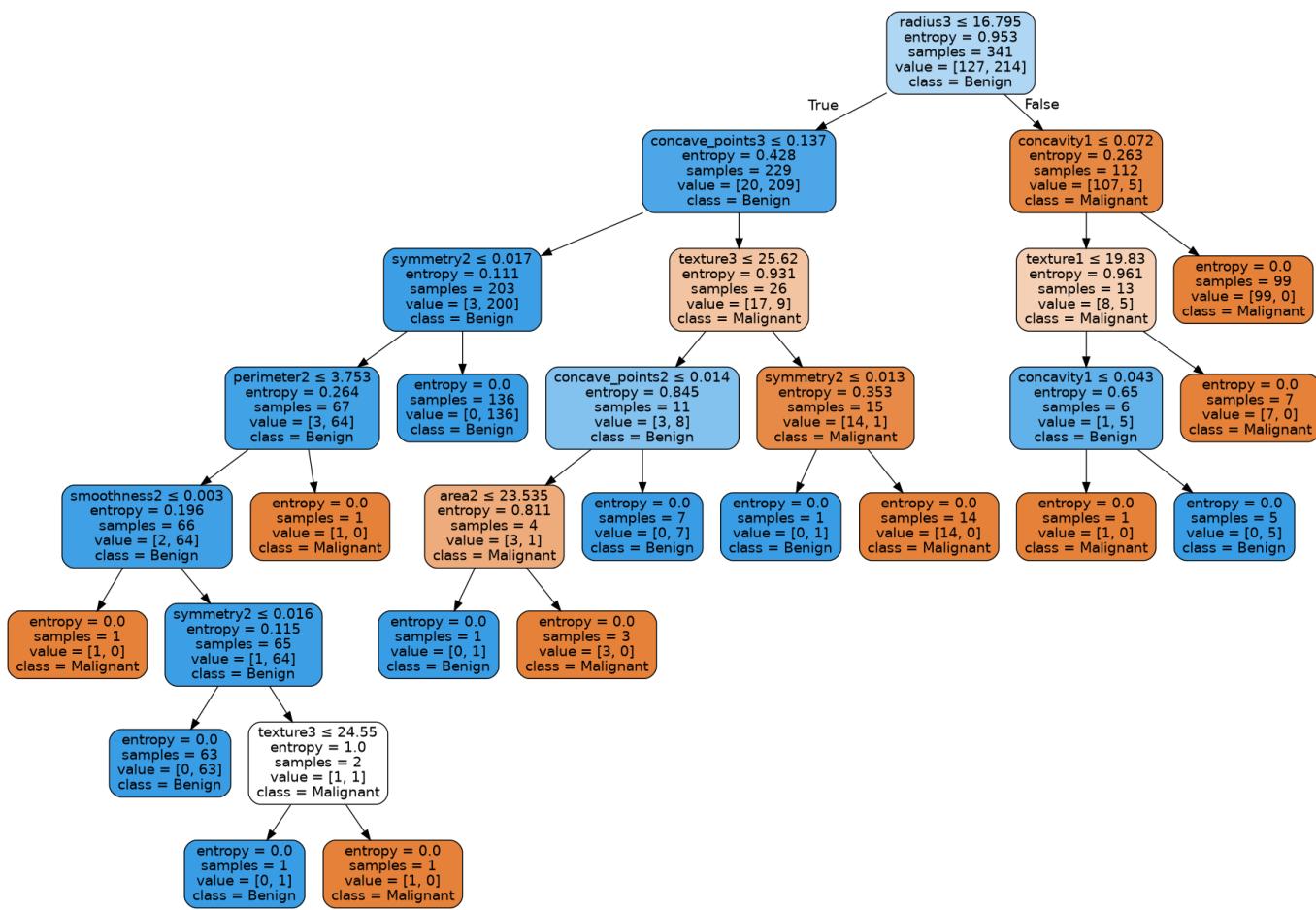
Bộ dữ liệu 40/60 (Train/Test)

Đường dẫn: Source\Cancer\Image4Datasets\DT_60_test.png



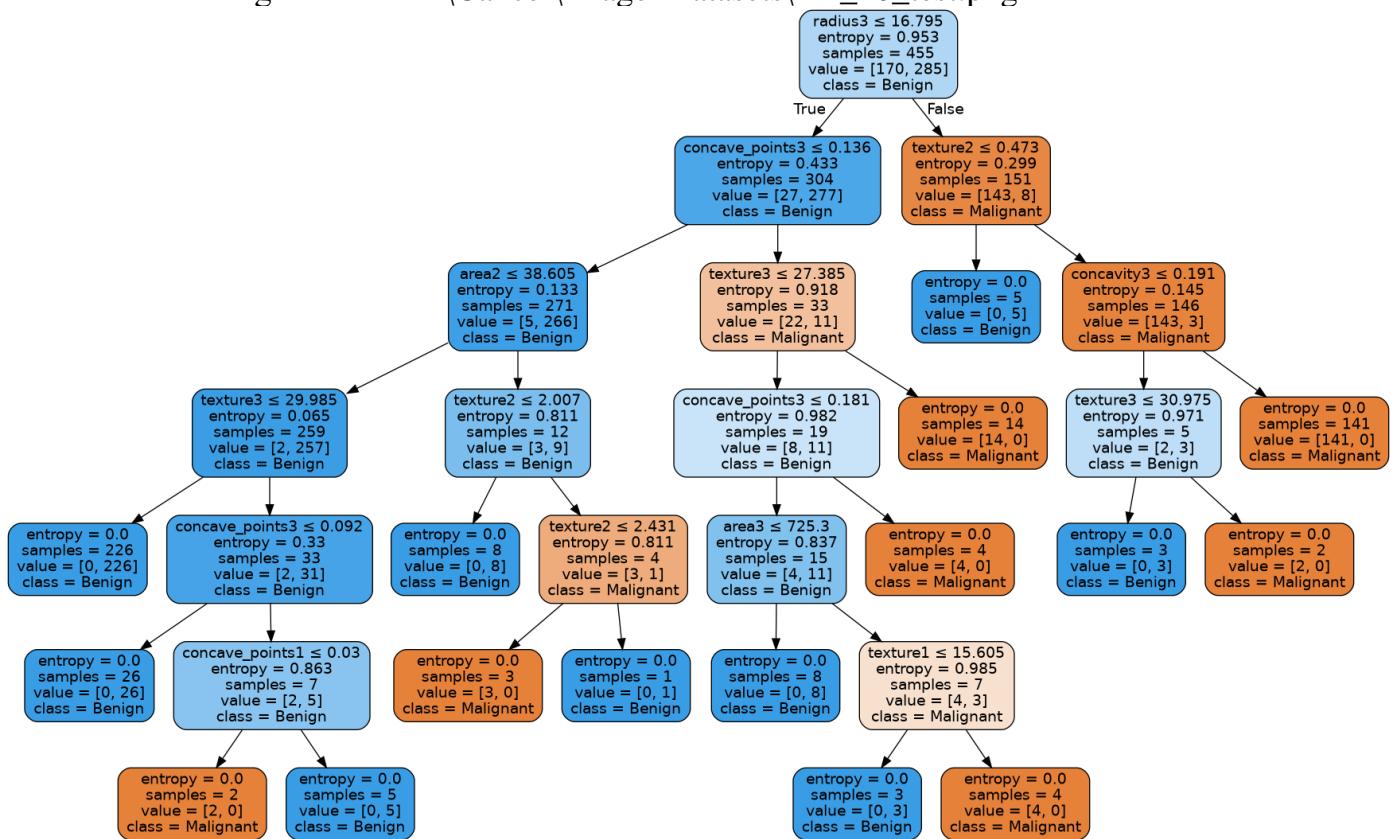
Bộ dữ liệu 60/40 (Train/ Test)

Đường dẫn: Source\Cancer\Image4Datasets\DT_40_test.png



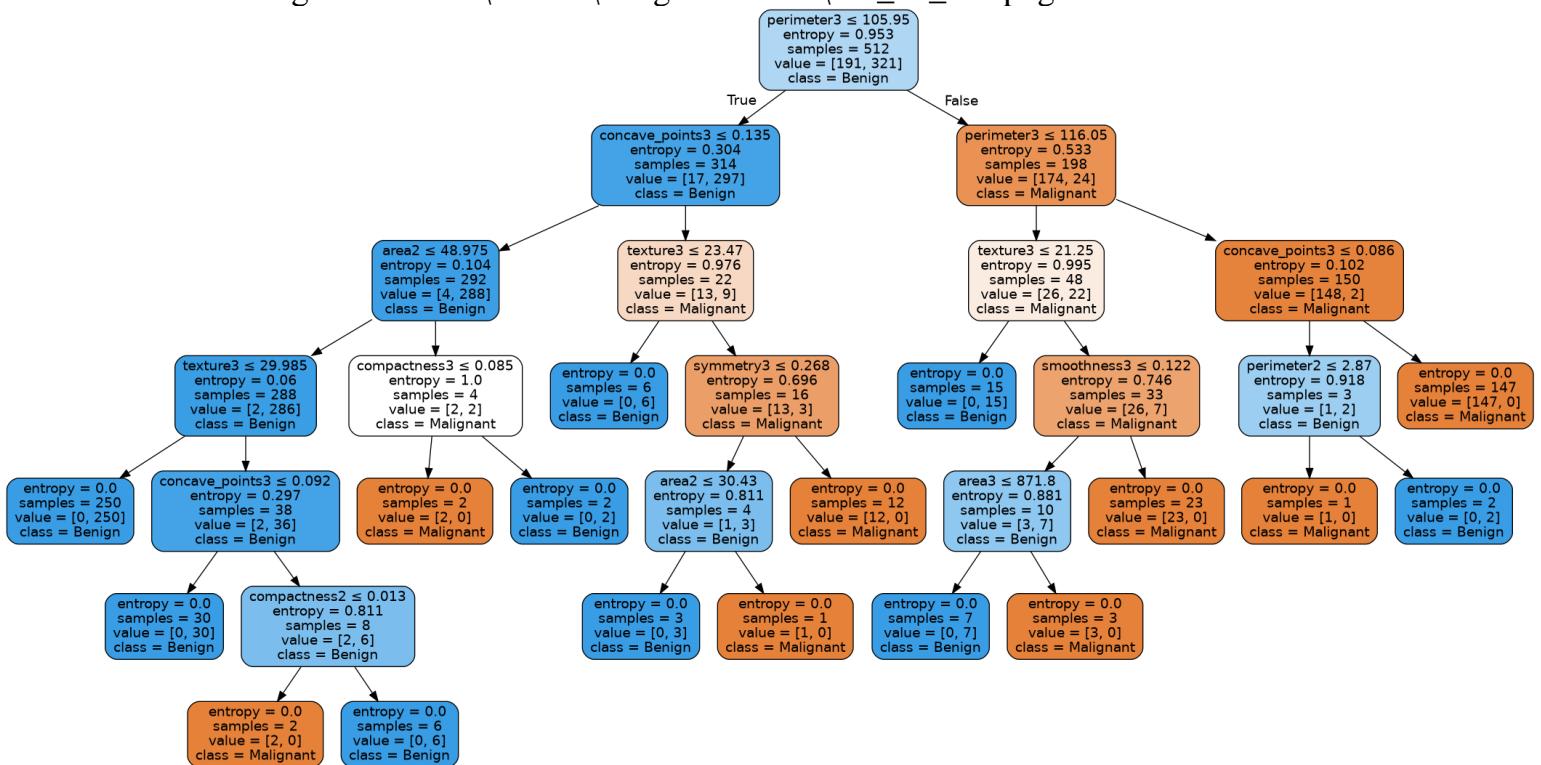
Bộ dữ liệu 80/20 (Train/Test)

Đường dẫn: Source\Cancer\Image4Datasets\DT_20_test.png



Bộ dữ liệu 90/10 (Train/Test)

Đường dẫn: Source\Cancer\Image4Datasets\DT_10_test.png



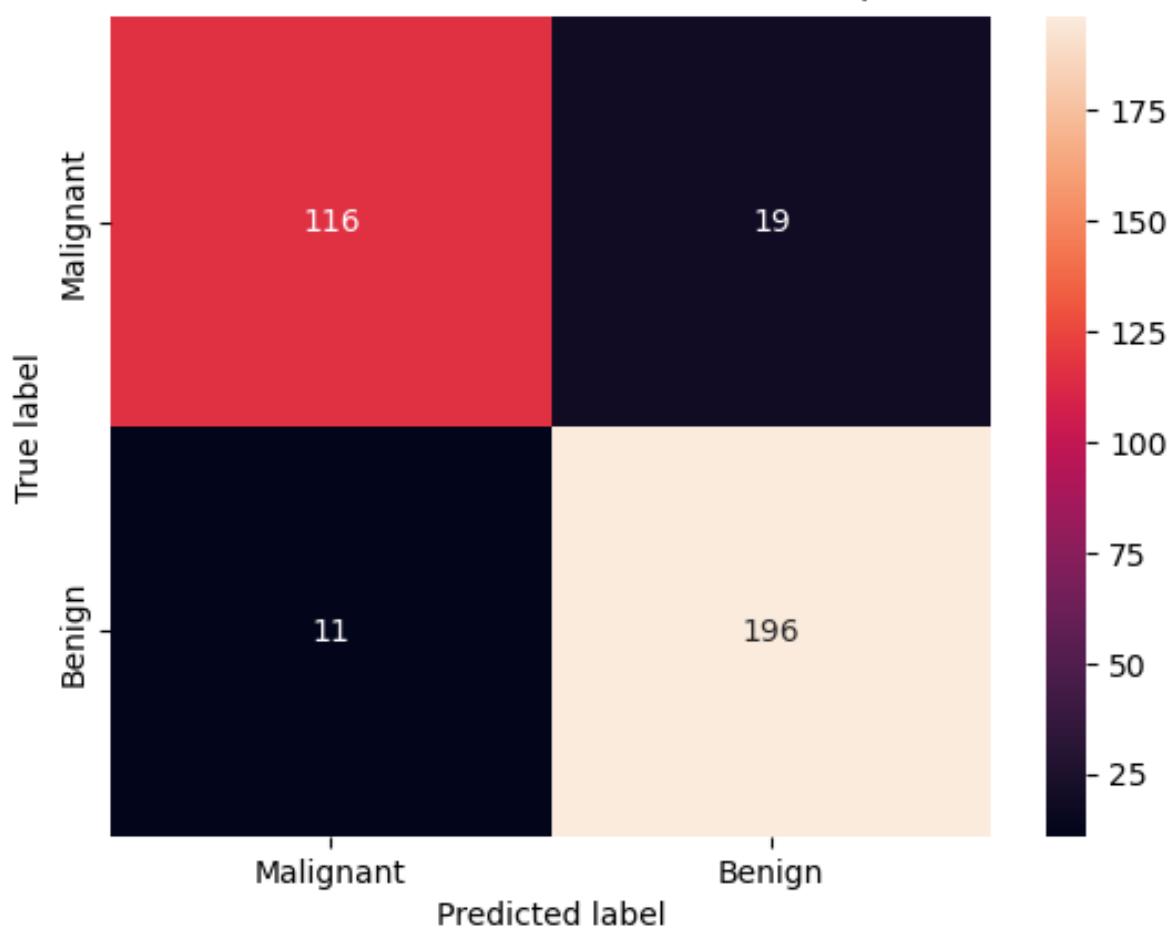
2.1.4 Đánh giá cây quyết định

Bộ dữ liệu 40/60 (Train/Test)

Classification Report:				
	precision	recall	f1-score	support
B	0.91	0.86	0.89	135
M	0.91	0.95	0.93	207
accuracy			0.91	342
macro avg	0.91	0.90	0.91	342
weighted avg	0.91	0.91	0.91	342

- Hiệu suất tổng thể: Mô hình đạt độ chính xác (Accuracy) là 91%, tức dự đoán đúng 91% số mẫu trong tập kiểm tra.
- Lớp B (Benign):
 - o Precision (0.91): Dự đoán lớp B chính xác 91%.
 - o Recall (0.86): Nhận diện được 86% các mẫu thực sự thuộc lớp B.
 - o F1-Score (0.89): Hiệu suất cân bằng giữa Precision và Recall ở mức khá tốt, nhưng Recall có thể được cải thiện.
- Lớp M (Malignant):
 - o Precision (0.91): Dự đoán lớp M chính xác 91%.
 - o Recall (0.95): Nhận diện gần như đầy đủ (95%) các mẫu thực sự thuộc lớp M.
 - o F1-Score (0.93): Hiệu suất cao, cho thấy mô hình hoạt động tốt với lớp này.
- Trung bình:
 - o Macro Average: Hiệu suất trung bình giữa hai lớp: Precision 0.91, Recall 0.90, F1-Score 0.91.
 - o Weighted Average: Hiệu suất trung bình có trọng số (dựa trên số mẫu mỗi lớp: Precision 0.91, Recall 0.91, F1-Score 0.91.

Decision Tree Classifier confusion matrix - 40/60 dataset



- Nhận xét: Với tỉ lệ tập train/test là 40/60:

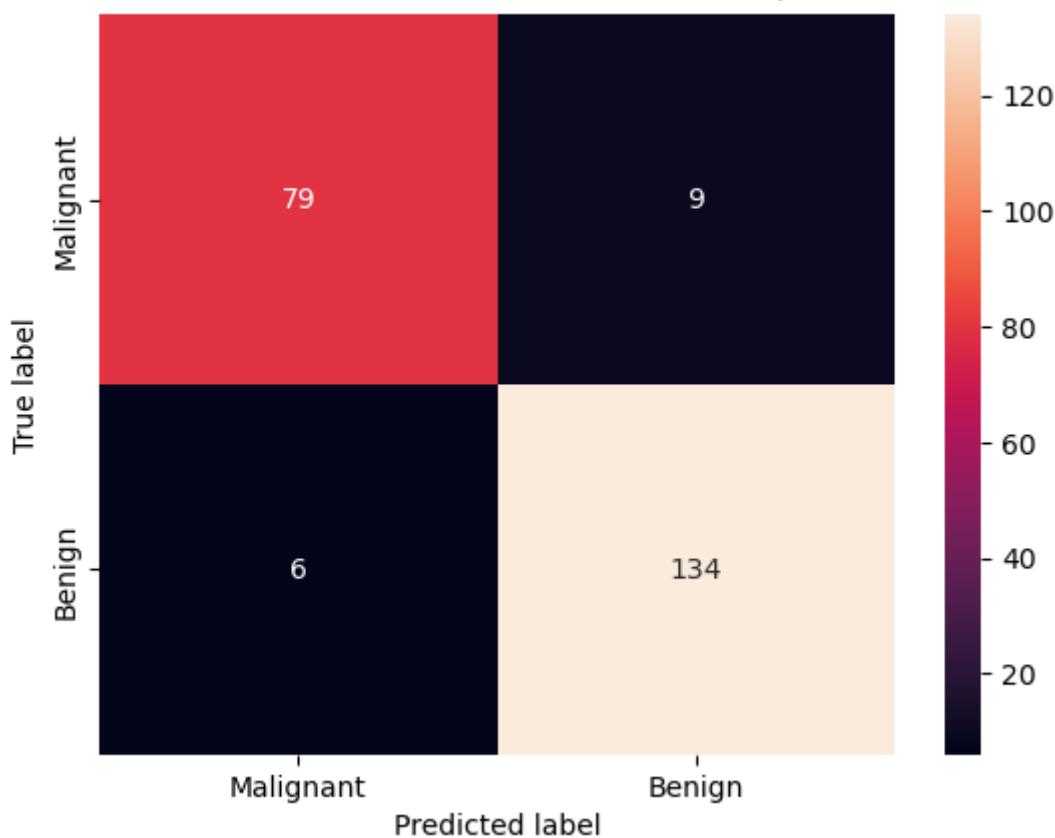
- Precision của cả 2 lớp là 0.91 nghĩa là 91% dự đoán thuộc 2 lớp là chính xác. Tỉ lệ này khá cao cho thấy mô hình ít dự đoán nhầm giữa các lớp.
- Recall của lớp Benign là 0.86 thấp hơn của lớp Malignant(M) là 0.95, cho thấy mô hình hơi ưu tiên nhận diện mẫu thuộc lớp Maglinant hơn lớp Benign.
- F1-Score và Accuracy của cả 2 lớp đều cao, cho thấy mô hình đạt hiệu quả tốt trong việc phân loại.

Bộ dữ liệu 60/40 (Train/Test)

Classification Report:				
	precision	recall	f1-score	support
B	0.93	0.90	0.91	88
M	0.94	0.96	0.95	140
accuracy			0.93	228
macro avg	0.93	0.93	0.93	228
weighted avg	0.93	0.93	0.93	228

- Hiệu suất tổng thể: Mô hình đạt độ chính xác (Accuracy) là 93%, tức dự đoán đúng 93% số mẫu trong tập kiểm tra.
- Lớp B (Benign):
 - o Precision (0.93): Dự đoán lớp B chính xác 93%.
 - o Recall (0.90): Nhận diện được 90% các mẫu thực sự thuộc lớp B.
 - o F1-Score (0.91): Hiệu suất cân bằng giữa Precision và Recall ở mức cao.
- Lớp M (Malignant):
 - o Precision (0.94): Dự đoán lớp M chính xác 94%.
 - o Recall (0.96): Nhận diện gần như đầy đủ (96%) các mẫu thực sự thuộc lớp M.
 - o F1-Score (0.95): Hiệu suất rất cao, cho thấy mô hình hoạt động tốt với lớp này.
- Trung bình:
 - o Macro Average: Hiệu suất trung bình giữa hai lớp: Precision 0.93, Recall 0.93, F1-Score 0.93.
 - o Weighted Average: Hiệu suất trung bình có trọng số (dựa trên số mẫu mỗi lớp): Precision 0.93, Recall 0.93, F1-Score 0.93.

Decision Tree Classifier confusion matrix - 60/40 dataset



- Nhận xét: Với tỉ lệ tập train/test là 60/40:

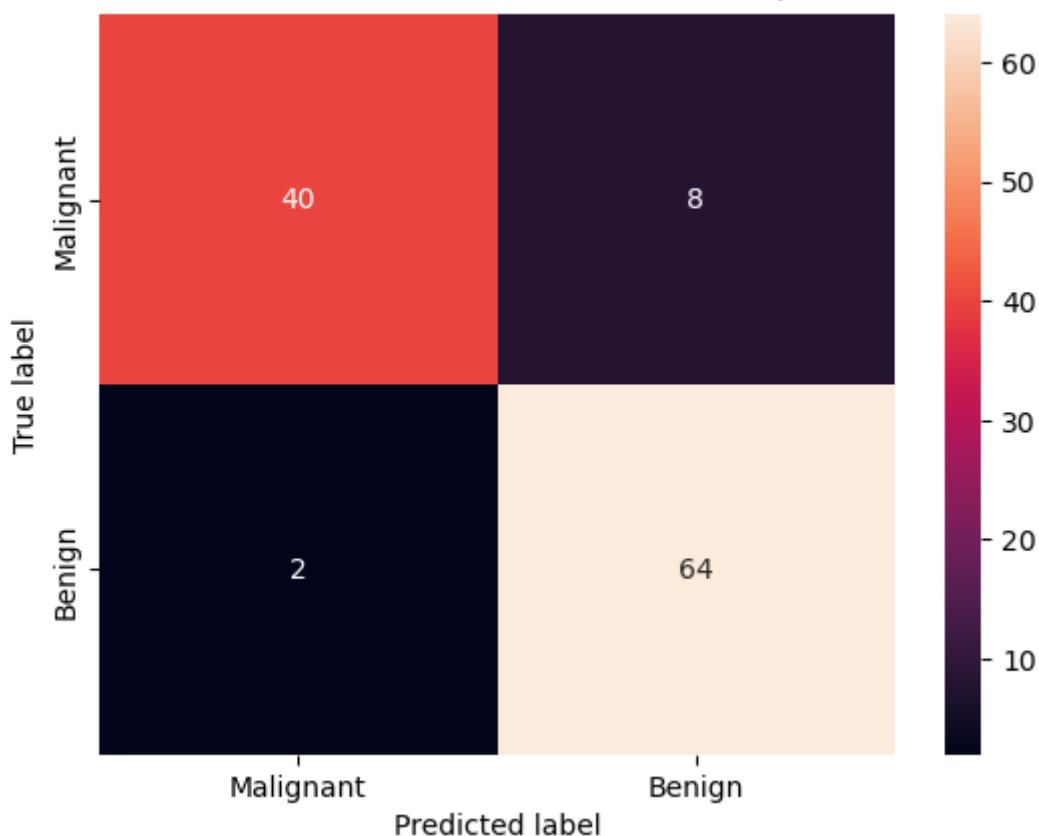
- Precision của lớp Malignant là 0.93 và của lớp Benign là 0.94. Tỉ lệ này cao hơn so với tập 40/60 train/test.
- Recall của lớp Benign là 0.90 thấp hơn của lớp Malignant(M) là 0.96, cho thấy mô hình hơi ưu tiên nhận diện mẫu thuộc lớp Malignant hơn lớp Benign. Mặt khác, khả năng nhận diện mẫu thuộc lớp Malignant cũng tăng lên.
- F1-Score và Accuracy của cả 2 lớp đều cao và cao hơn so với tập 40/60 train/test, cho thấy mô hình đạt hiệu quả phân loại tốt hơn với tỉ lệ train/test trước đó.

Bộ dữ liệu 80/20 (Train/Test)

Classification Report:					
	precision	recall	f1-score	support	
B	0.95	0.83	0.89	48	
M	0.89	0.97	0.93	66	
accuracy			0.91	114	
macro avg	0.92	0.90	0.91	114	
weighted avg	0.92	0.91	0.91	114	

- Hiệu suất tổng thể: Mô hình đạt độ chính xác (Accuracy) là 91%, tức dự đoán đúng 91% số mẫu trong tập kiểm tra.
- Lớp B (Benign):
 - o Precision (0.95): Dự đoán lớp B chính xác 95%.
 - o Recall (0.83): Nhận diện được 83% các mẫu thực sự thuộc lớp B.
 - o F1-Score (0.89): Hiệu suất cân bằng giữa Precision và Recall ở mức tốt.
- Lớp M (Malignant):
 - o Precision (0.89): Dự đoán lớp M chính xác 89%.
 - o Recall (0.97): Nhận diện gần như đầy đủ (97%) các mẫu thực sự thuộc lớp M.
 - o F1-Score (0.93): Hiệu suất cao, cho thấy mô hình hoạt động tốt với lớp này.
- Trung bình:
 - o Macro Average: Hiệu suất trung bình giữa hai lớp: Precision 0.92, Recall 0.90, F1-Score 0.91.
 - o Weighted Average: Hiệu suất trung bình có trọng số (dựa trên số mẫu mỗi lớp): Precision 0.92, Recall 0.91, F1-Score 0.91.

Decision Tree Classifier confusion matrix - 80/20 dataset



- Nhận xét: Với tỉ lệ tập train/test là 80/20:

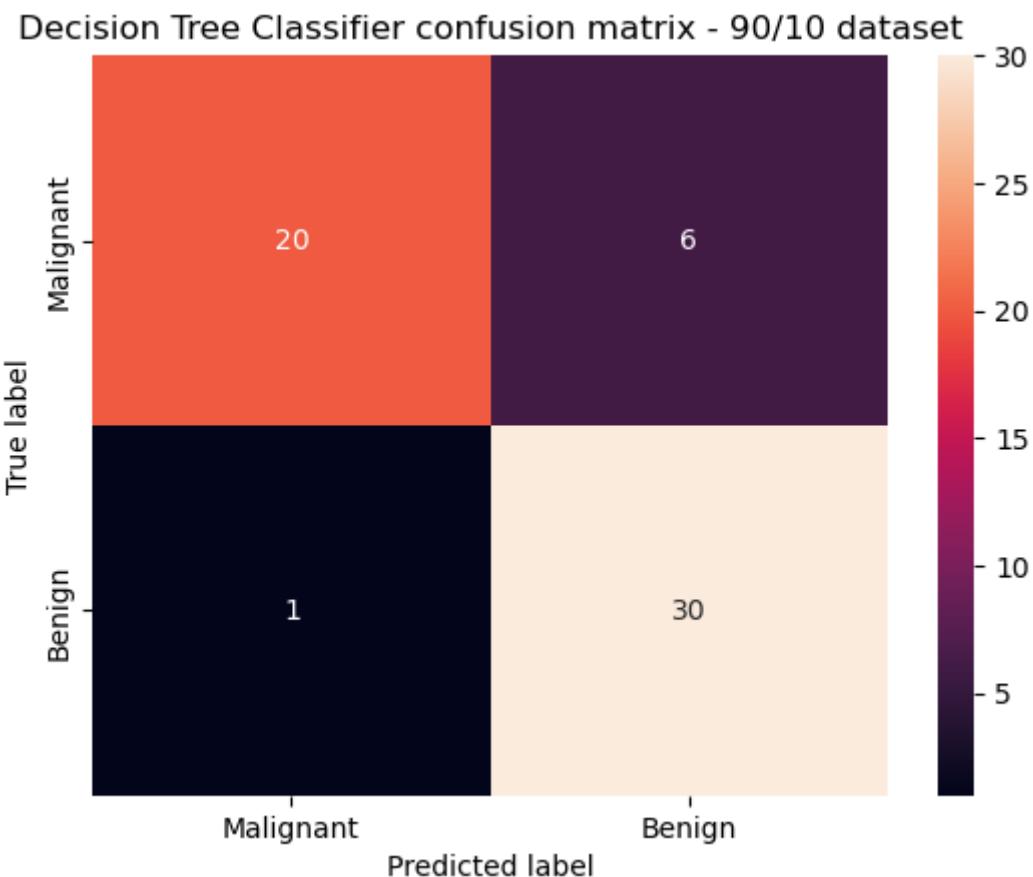
- Precision của lớp Malignant là 0.95 và của lớp Benign là 0.89. Tỉ lệ này đã có sự thiên lệch nhiều so với 2 tập dữ liệu trước đó.
- Recall của lớp Benign là 0.83 thấp hơn nhiều so với lớp Malignant(M) là 0.97, cho thấy mô hình lúc này ưu tiên nhận diện mẫu thuộc lớp Malignant hơn lớp Benign. Khả năng nhận diện mẫu thuộc lớp Malignant tăng lên 1 chút, nhưng khả năng nhận diện mẫu của lớp Benign giảm đáng kể.
- F1-Score và Accuracy của cả 2 lớp đều cao nhưng vẫn thấp hơn so với tập 40/60 train/test, cho thấy mô hình đạt hiệu quả phân loại tốt.

Bộ dữ liệu 90/10 (Train/Test)

Classification Report:

	precision	recall	f1-score	support
B	0.95	0.77	0.85	26
M	0.83	0.97	0.90	31
accuracy			0.88	57
macro avg	0.89	0.87	0.87	57
weighted avg	0.89	0.88	0.88	57

- Hiệu suất tổng thể: Mô hình đạt độ chính xác (Accuracy) là 88%, tức dự đoán đúng 88% số mẫu trong tập kiểm tra
- Lớp B (Benign):
 - o Precision (0.95): Dự đoán lớp B chính xác 95%.
 - o Recall (0.77): Nhận diện được 77% các mẫu thực sự thuộc lớp B.
 - o F1-Score (0.85): Hiệu suất cân bằng giữa Precision và Recall ở mức khá tốt.
- Lớp M (Malignant):
 - o Precision (0.83): Dự đoán lớp M chính xác 83%.
 - o Recall (0.97): Nhận diện gần như đầy đủ (97%) các mẫu thực sự thuộc lớp M.
 - o F1-Score (0.90): Hiệu suất cao, cho thấy mô hình hoạt động tốt với lớp này.
- Trung bình:
 - Macro Average: Hiệu suất trung bình giữa hai lớp: Precision 0.89, Recall 0.87, F1-Score 0.87.
 - Weighted Average: Hiệu suất trung bình có trọng số (dựa trên số mẫu mỗi lớp): Precision 0.89, Recall 0.88, F1-Score 0.88.

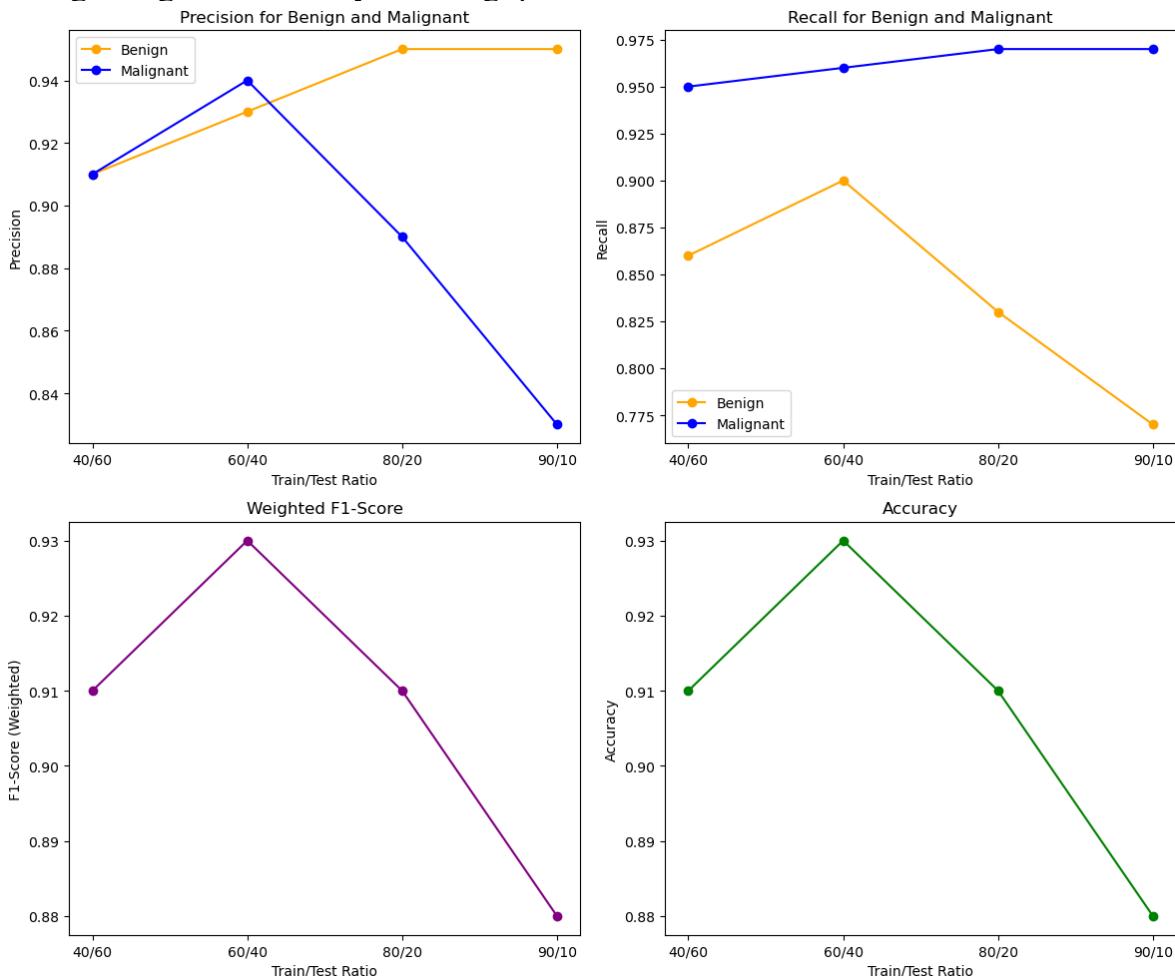


- Nhận xét: Với tỉ lệ tập train/test là 90/10:

- Precision của lớp Malignant là 0.95 và của lớp Benign là 0.83. Tỉ lệ này có sự thiên lệch nhiều hơn so với 3 tập dữ liệu trước đó.
- Recall của lớp Benign là 0.77 thấp hơn nhiều so với lớp Malignant(M) là 0.97, cho thấy mô hình lúc này ưu tiên nhận diện mẫu thuộc lớp Malignant hơn lớp Benign. Khả năng nhận diện mẫu thuộc lớp Malignant không đổi, nhưng khả năng nhận diện mẫu của lớp Benign giảm đáng kể.
- F1-Score và Accuracy của cả 2 lớp đều cao nhưng thấp nhất trong 4 cách chia, cho thấy mô hình đạt hiệu quả phân loại ở mức khá tốt.

Insight

Với dữ liệu quan sát ở trên, nhóm chọn ra các thông tin là precision, recall và accuracy là những thông tin có sự thay đổi đáng quan sát

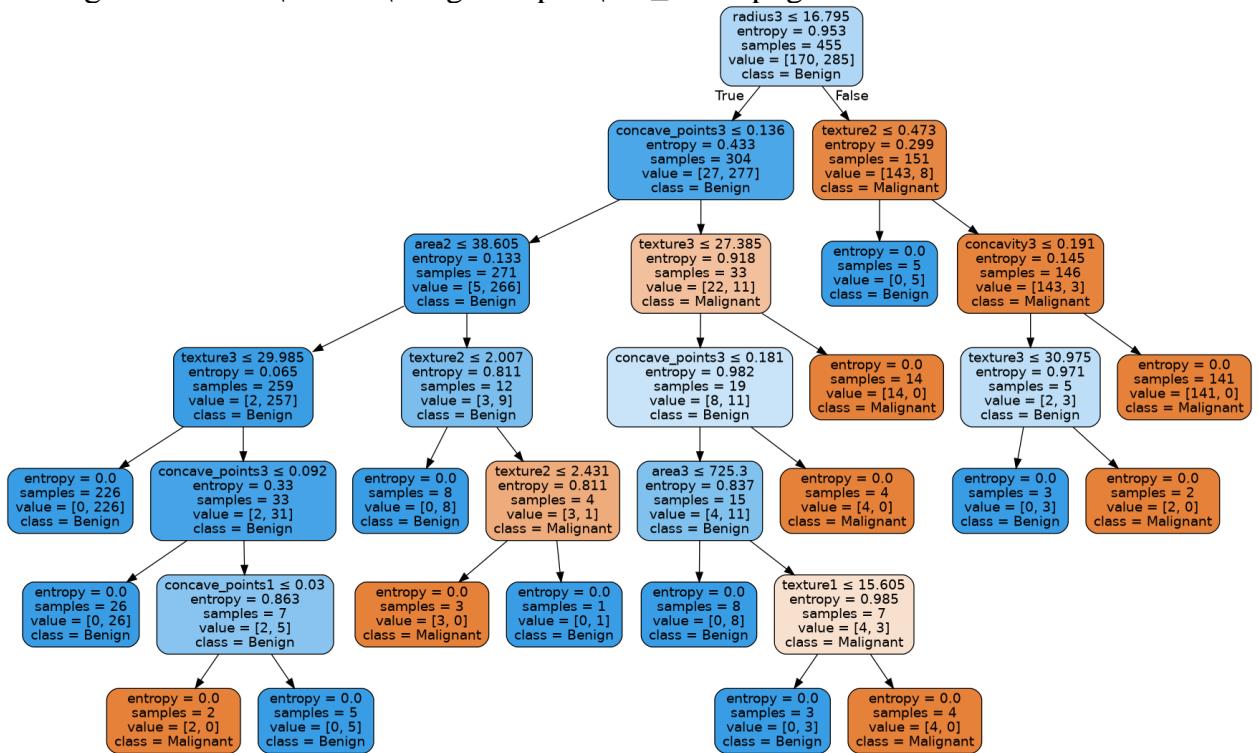


- Với phân lớp Benign, có thể thấy rằng khi tỉ lệ tập test giảm từ 0.6 xuống 0.2 thì tỉ lệ dự đoán chính xác mẫu của lớp này tăng, khi giảm tới 0.1 thì có sự giảm nhẹ. Có thể thấy điều này qua sự tăng mạnh của precision và recall. Điều này cho thấy là nhìn chung thì tỉ lệ dự đoán chính xác lớp Benign tăng khi tỉ lệ tập test giảm dần tới 0.2.
- Với phân lớp Malignant, có thể thấy rằng precision và recall của Malignant có sự tăng trưởng rõ rệt từ 40/60 lên 60/40 nhưng lại giảm xuống ở các tỷ lệ tiếp theo. Điều này cho thấy rằng với tỉ lệ tập test nhỏ hơn 0.4 và giảm dần, thì tỉ lệ dự đoán chính xác mẫu của lớp Malignant giảm dần.
- Tại tỉ lệ train/test là 60/40, accuracy và weighted F1-score là cao nhất cho tập test và giảm dần khi tỉ lệ test giảm dần. Điều này cho thấy hiệu suất giảm khi huấn luyện trên nhiều dữ liệu.
- Nhìn chung, mô hình có xu hướng phân loại Benign chính xác hơn. Dữ liệu của các lớp trong tập test cũng theo phân phối của dữ liệu gốc - là phân phối không đều với tỉ lệ Benign/Malignant xấp xỉ 3/2, nên mô hình dường như ưu tiên lớp Benign và càng thiên vị khi tỉ lệ tập test nhỏ hơn. Và mặc dù weighted F1-score giảm khi giảm tỉ lệ test, nhưng sự thay đổi này vẫn không lớn. Điều này cho thấy mô hình vẫn đang duy trì hiệu suất khá ổn định.

2.1.5 Phân tích độ sâu và độ chính xác

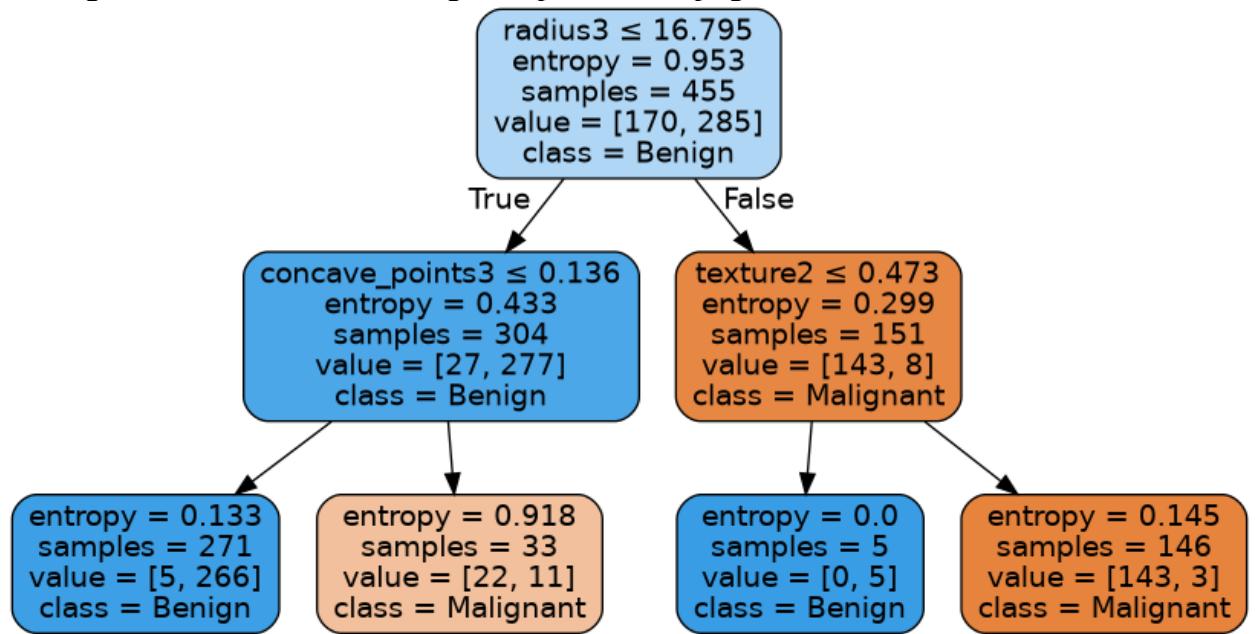
Max depth = None (Không có độ sâu giới hạn)

Đường dẫn: Source\Cancer\Image4Depths\DT_None.png



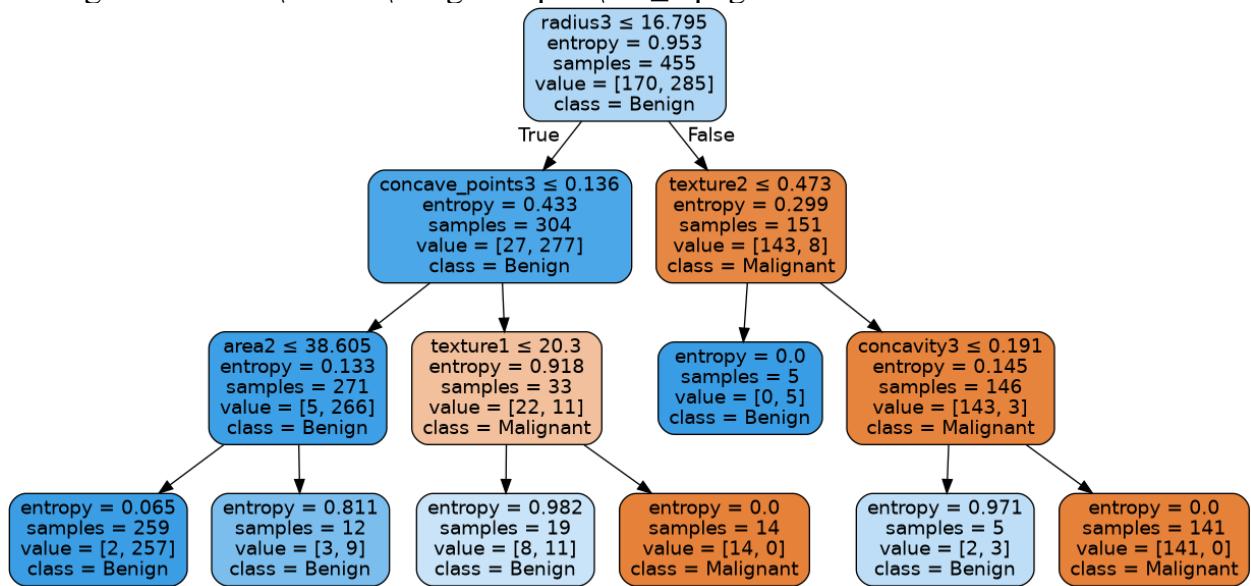
Max depth = 2

Đường dẫn: Source\Cancer\Image4Depths\DT_2.png



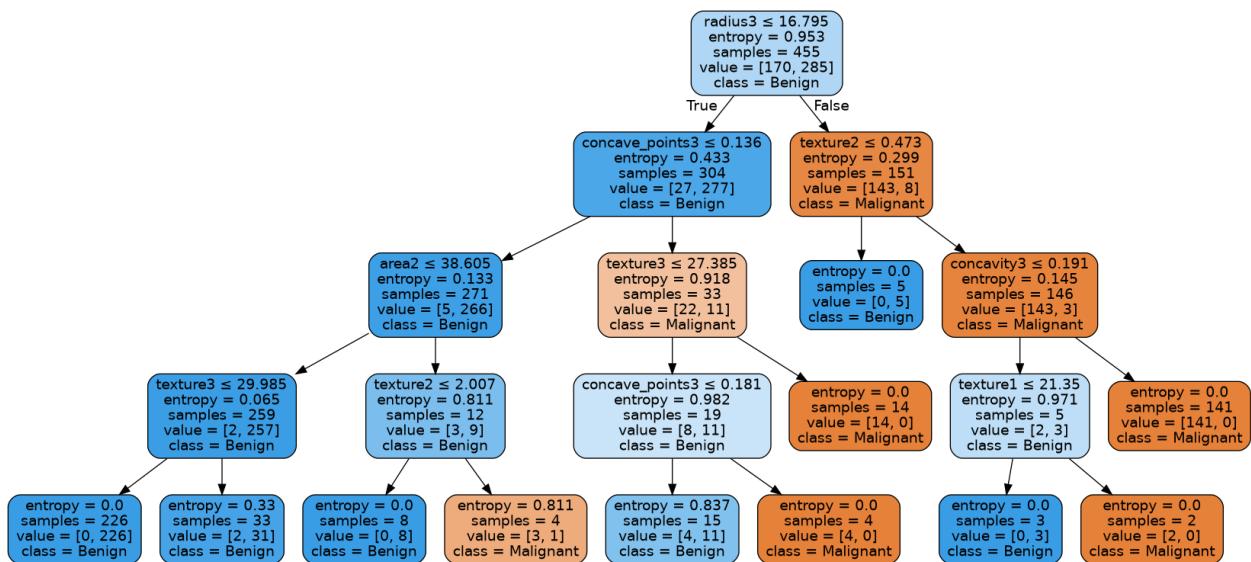
Max depth = 3

Đường dẫn: Source\Cancer\Image4Depths\DT_3.png



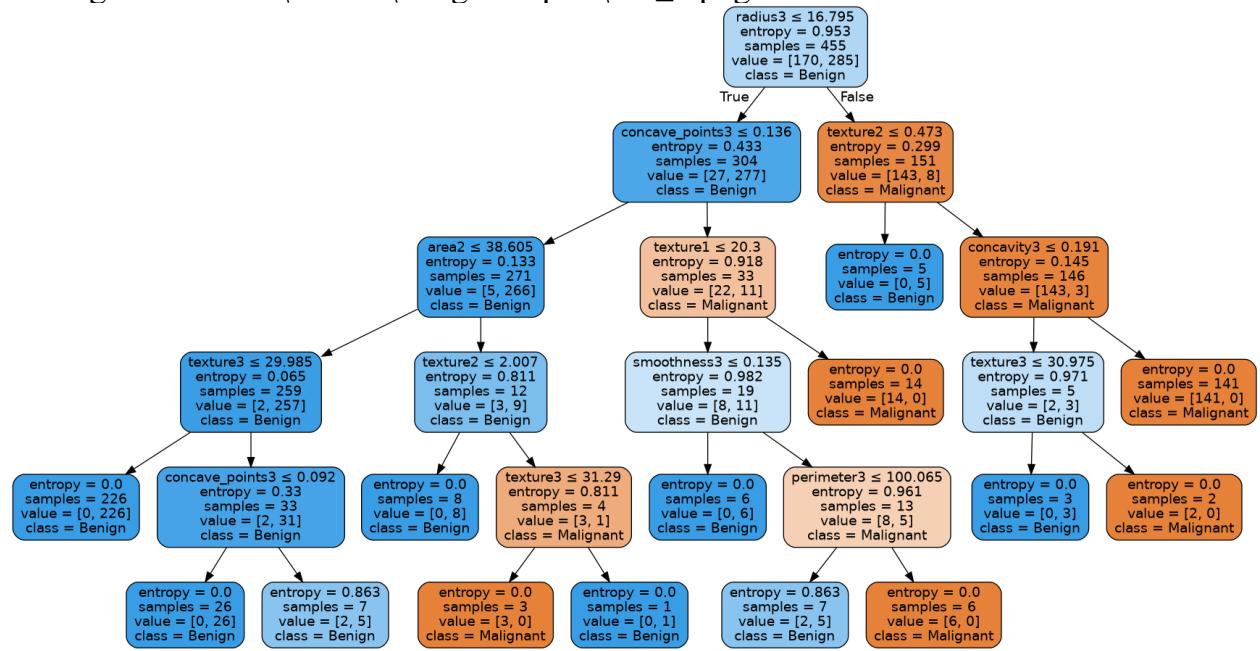
Max depth = 4

Đường dẫn: Source\Cancer\Image4Depths\DT_4.png



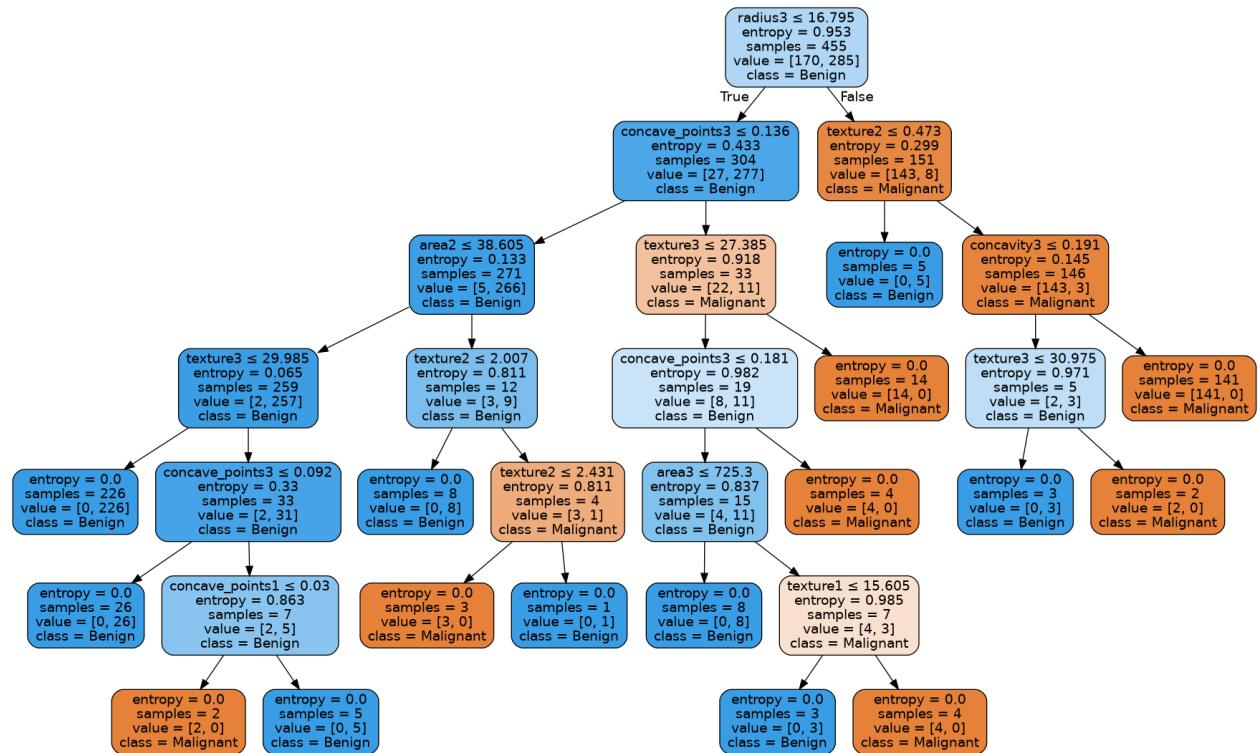
Max depth = 5

Đường dẫn: Source\Cancer\Image4Depths\DT_5.png



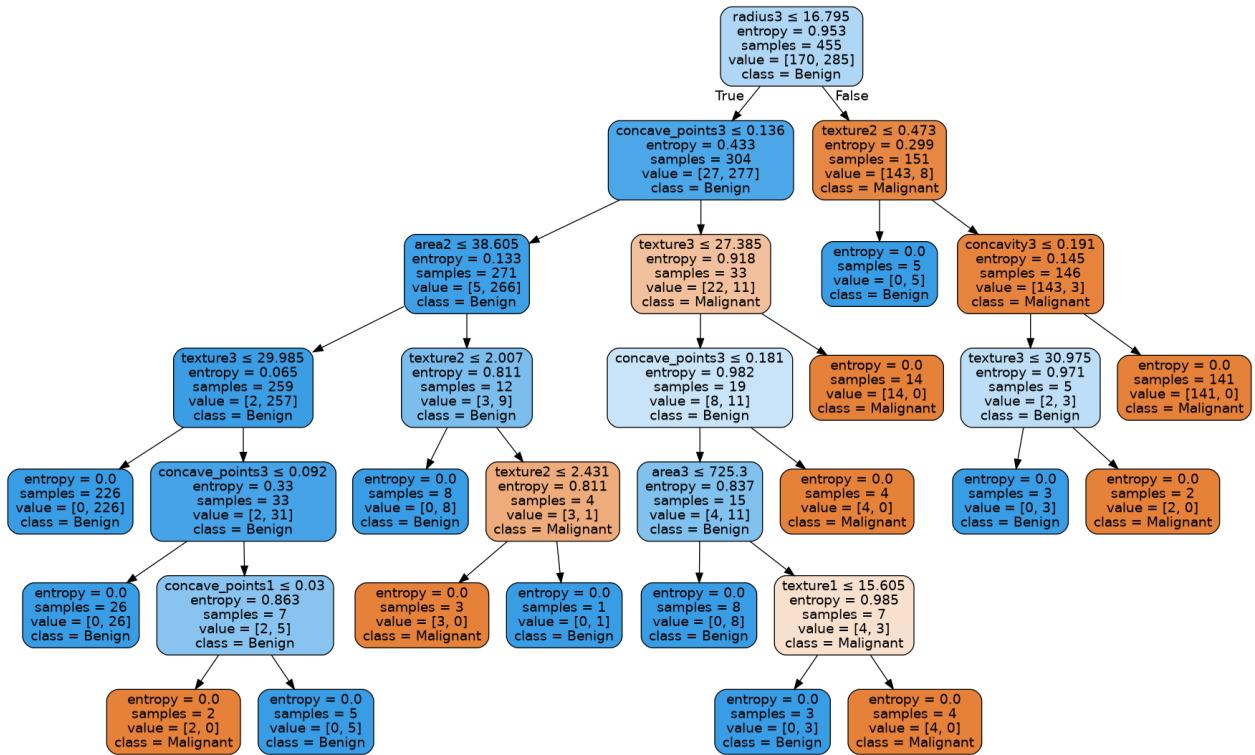
Max depth = 6

Đường dẫn: Source\Cancer\Image4Depths\DT_6.png

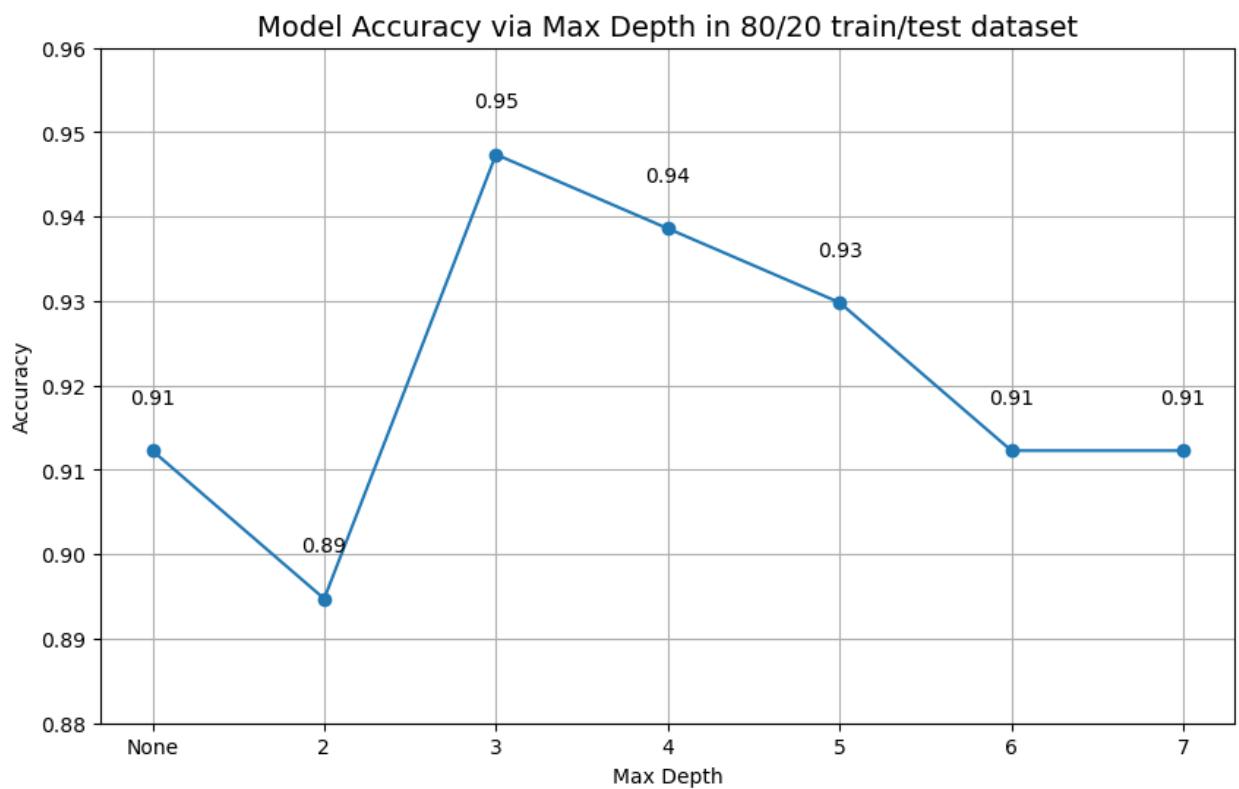


Max depth = 7

Đường dẫn: Source\Cancer\Image4Depths\DT_7.png



So sánh các độ sâu



Insight:

- Từ kết quả hình trên, có thể thấy rằng chiều sâu của decision tree cho tập 80/20 lớn nhất là 6. Vì tại 7 và None (mặc định) thì độ chính xác đều là 0,91.
- Độ sâu max_depth bằng 3 là tối ưu nhất cho tập dữ liệu 80/20 train/test vì chúng cho kết quả độ chính xác cao nhất là 0,95.
- Khi độ sâu quá lớn, tức là max_depth lớn hơn 3 và tăng dần lên, độ chính xác giảm dần, cho thấy cây có thể bị overfitting khi học quá nhiều và làm cây trở nên quá phức tạp.
- Khi độ sâu quá nhỏ, tức là khi max_depth = 2, độ chính xác khá thấp, cho thấy cây không học đủ để phân biệt các mẫu của 2 lớp dẫn đến không mô hình hóa được các quan hệ giữa các đặc điểm.

2.2 Wine Quality Dataset

2.2.1 Chuẩn bị dữ liệu

- Bộ dữ liệu sử dụng trong dự án này là Wine Quality Dataset, được tải bằng hàm `fetch_ucirepo` từ module `ucimlrepo` với ID = 186. Bộ dữ liệu chứa thông tin về các loại rượu vang, bao gồm:
 - o Số đặc trưng (Features): 11, bao gồm các chỉ số hóa học như độ axit, hàm lượng cồn, và độ pH.
 - o Số mẫu (Samples): 4898.
 - o Đầu ra (Targets): Chất lượng rượu được đánh giá theo thang điểm từ 0 đến 10.
- Từ dữ liệu đầu vào chia thành 3 nhóm
 - o Low quality : nhóm 0 (lớp 0-4)
 - o Standard quality: nhóm 1 (lớp 5-6)
 - o High quality : nhóm 2 (lớp 7-10)

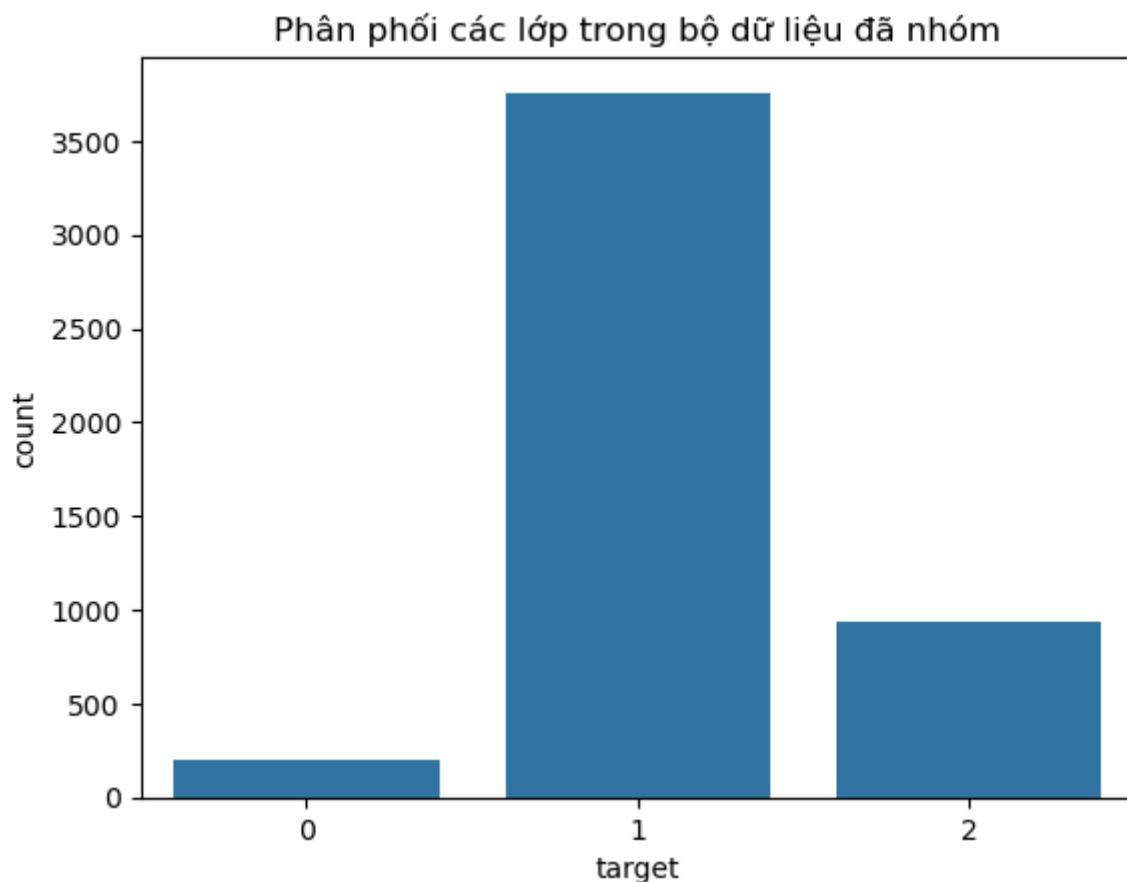
Features shape: (4898, 11)

Targets shape: (4898, 1)

```
fixed_acidity  volatile_acidity  citric_acid  residual_sugar  chlorides  \
0            7.4              0.70          0.00            1.9       0.076
1            7.8              0.88          0.00            2.6       0.098
2            7.8              0.76          0.04            2.3       0.092
3           11.2              0.28          0.56            1.9       0.075
4            7.4              0.70          0.00            1.9       0.076

free_sulfur_dioxide  total_sulfur_dioxide  density  pH  sulphates  \
0            11.0              34.0  0.9978  3.51       0.56
1            25.0              67.0  0.9968  3.20       0.68
2            15.0              54.0  0.9970  3.26       0.65
3            17.0              60.0  0.9980  3.16       0.58
4            11.0              34.0  0.9978  3.51       0.56

alcohol
0            9.4
1            9.8
2            9.8
3            9.8
4            9.4
```

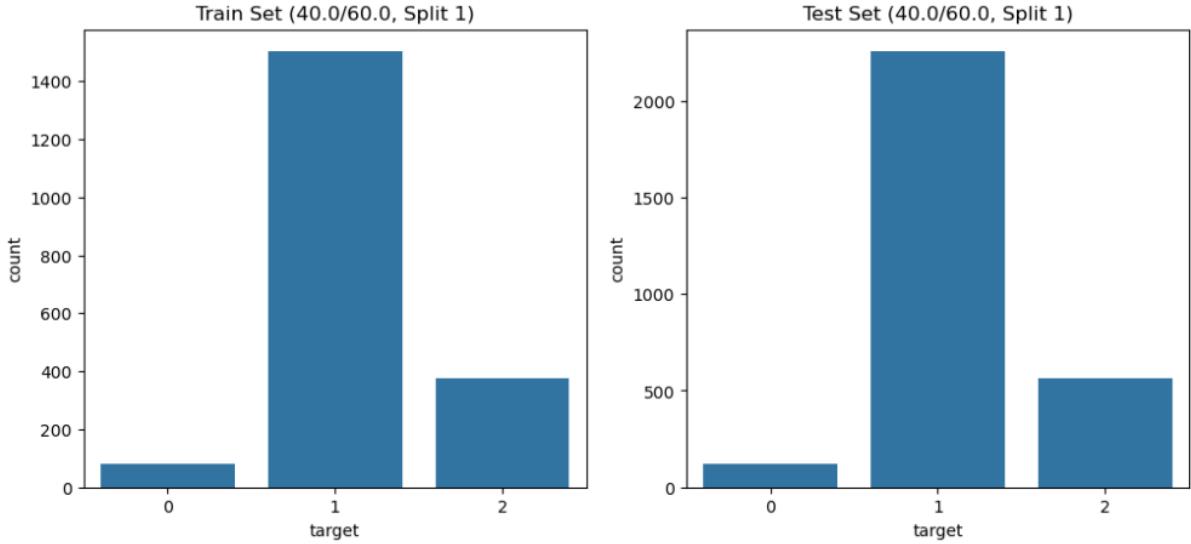


- Dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo các tỷ lệ khác nhau (40%, 60%, 80%, và 90%) để đánh giá hiệu suất của mô hình dưới các kích thước tập huấn luyện khác nhau. Dùng StratifiedShuffleSplit để tạo tập huấn luyện và kiểm tra, đảm bảo phân phối lớp giữa các tập giống như dữ liệu ban đầu.

- Trực quan hóa tỷ lệ 40/60 (Train/Test)

Tỷ lệ 40.0/60.0, Lần chia thứ 1:
Phân phối lớp trong tập huấn luyện:

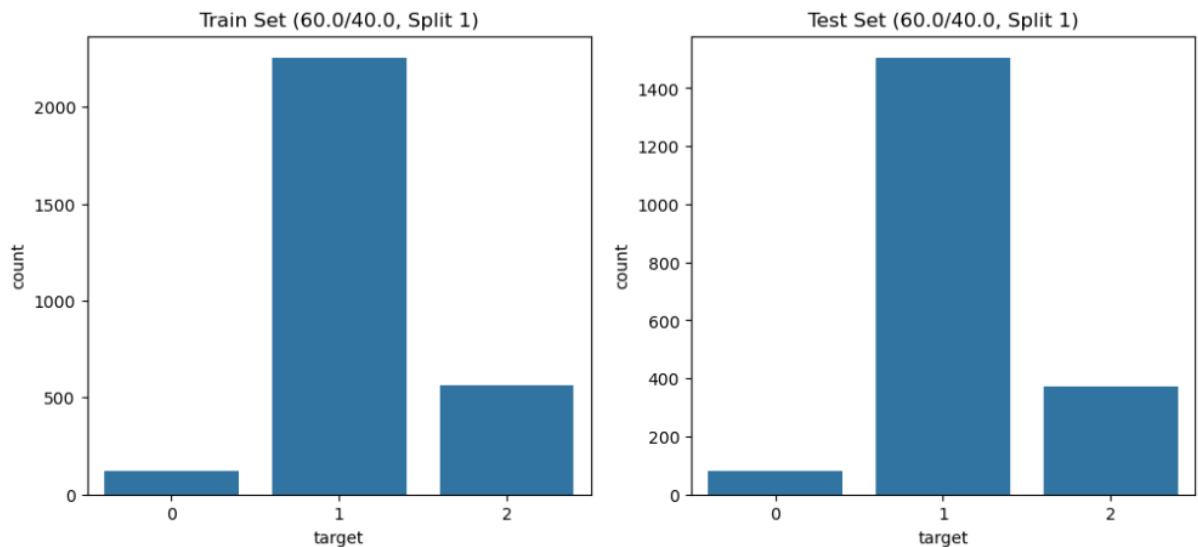
target
1 0.767228
2 0.190914
0 0.041858
Name: proportion, dtype: float64
Phân phối lớp trong tập kiểm tra:
target
1 0.767268
2 0.191222
0 0.041511
Name: proportion, dtype: float64



- Trực quan hóa tỷ lệ 60/40 (Train/Test)

Tỷ lệ 60.0/40.0, Lần chia thứ 1:
Phân phối lớp trong tập huấn luyện:

target
1 0.767189
2 0.191287
0 0.041525
Name: proportion, dtype: float64
Phân phối lớp trong tập kiểm tra:
target
1 0.767347
2 0.190816
0 0.041837
Name: proportion, dtype: float64



- Trực quan hóa tỷ lệ 80/20 (Train/Test)

Tỷ lệ 80.0/20.0, Lần chia thứ 1:

Phân phối lớp trong tập huấn luyện:

target

1 0.767228
2 0.191169
0 0.041603

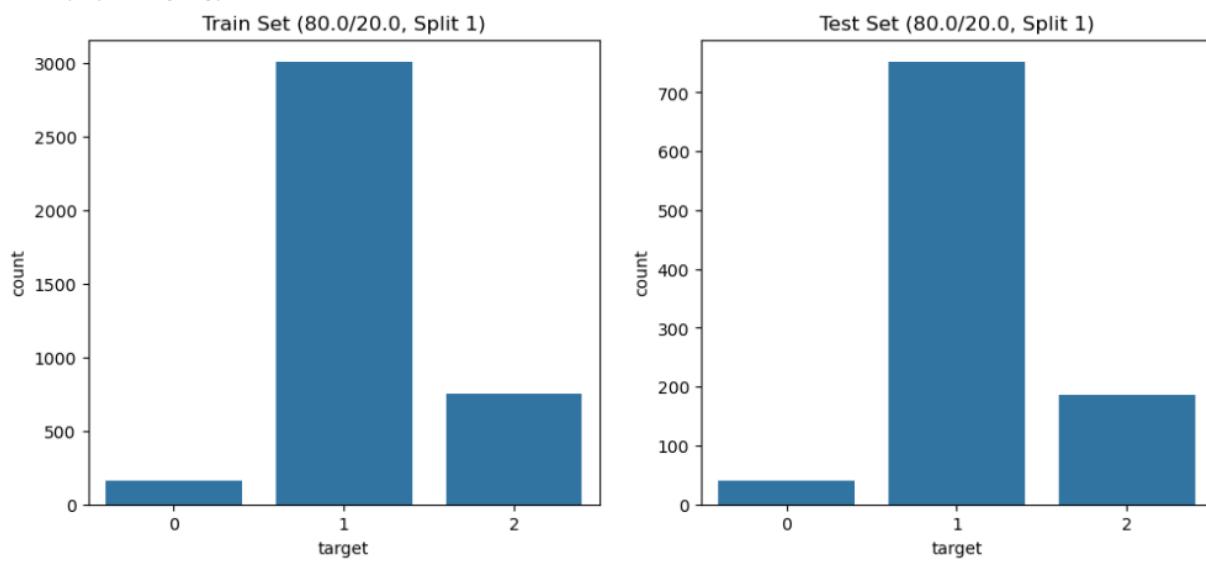
Name: proportion, dtype: float64

Phân phối lớp trong tập kiểm tra:

target

1 0.767347
2 0.190816
0 0.041837

Name: proportion, dtype: float64



- Trực quan hóa tỷ lệ 90/10 (Train/Test)

Tỷ lệ 90.0/10.0, Lần chia thứ 1:

Phân phối lớp trong tập huấn luyện:

target

1 0.767241
2 0.191016
0 0.041742

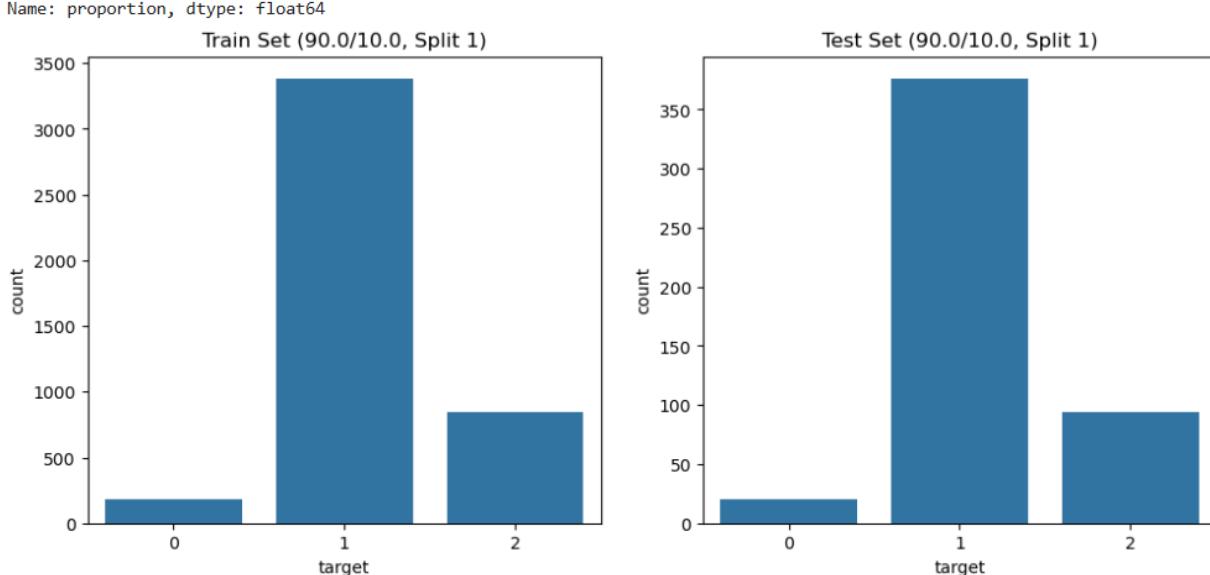
Name: proportion, dtype: float64

Phân phối lớp trong tập kiểm tra:

target

1 0.767347
2 0.191837
0 0.040816

Name: proportion, dtype: float64



Nhận xét

- Phân phối gốc (Original Dataset): Tỷ lệ các lớp
 - o Lớp 1 (Standard quality): 76.7%.
 - o Lớp 2 (High quality): 19.1%.
 - o Lớp 0 (Low quality): 4.2%
 - o Tập dữ liệu ban đầu không cân bằng, với lớp Standard quality chiếm phần lớn và hai lớp còn lại chiếm tỷ lệ nhỏ hơn (đặc biệt lớp Low quality chỉ 4.2%). Chênh lệch này khá lớn
- Phân phối các tập con: Tỷ lệ các lớp trong tập huấn luyện và kiểm tra
 - o Ở tất cả các tỷ lệ phân chia (40/60, 60/40, 80/20, 90/10), phân phối lớp trong cả tập huấn luyện và tập kiểm tra gần như giữ nguyên so với tập dữ liệu gốc

2.2.2 Cài đặt

- Cây quyết định được xây dựng bằng DecisionTreeClassifier từ scikit-learn, với tiêu chí phân chia là entropy.
- Dữ liệu được truyền vào mô hình thông qua hàm fit, và nhãn được dự đoán bằng hàm predict.
- Xuất mã DOT bằng export_graphviz: Lưu cấu trúc cây với đầy đủ thông tin nút phân chia và nút lá.
- Sử dụng graphviz: Chuyển mã DOT thành hình ảnh cây với các nút làm tròn và tô màu theo tỷ lệ lớp.
- Nội dung hiển thị trên cây:
 - o Nút phân chia: đặc trưng, ngưỡng phân chia tại của đặc trưng, entropy, số lượng mẫu, phân phối lớp, nhãn.
- Nút lá: entropy, số lượng mẫu, phân phối lớp, nhãn.

2.2.3 Trực quan hóa cây quyết định

Vì các ảnh cây quyết định của dataset này có rất nhiều nhánh và phân nhiều tầng nên việc đưa ảnh vào báo cáo không thể thấy được rõ ràng. Nhóm em xin phép chỉ ghi đường dẫn của file ảnh trong folder a.

Bộ dữ liệu 40/60 (Train/Test)

Đường dẫn: Source\Wine\Image4Datasets\decision_tree_40_60_split_1.png

Bộ dữ liệu 60/40 (Train/ Test)

Đường dẫn: Source\Wine\Image4Datasets\decision_tree_60_40_split_1.png

Bộ dữ liệu 80/20 (Train/Test)

Đường dẫn: Source\Wine\Image4Datasets\decision_tree_80_20_split_1.png

Bộ dữ liệu 90/10 (Train/Test)

Đường dẫn: Source\Wine\Image4Datasets\decision_tree_90_10_split_1.png

2.2.4 Đánh giá cây quyết định

Bộ dữ liệu 40/60 (Train/Test)

Train/Test Split: 40/60, Lần chia thứ 1

Classification Report:

	precision	recall	f1-score	support
Low quality	0.18	0.15	0.16	122
Standard quality	0.84	0.85	0.84	2255
High quality	0.51	0.52	0.51	562
accuracy			0.75	2939
macro avg	0.51	0.50	0.51	2939
weighted avg	0.75	0.75	0.75	2939

- Hiệu suất tổng thể: Mô hình đạt độ chính xác (Accuracy) là 75%, tức dự đoán đúng 75% số mẫu trong tập kiểm tra.
- Lớp Low quality:
 - o Precision (0.18): Chỉ 18% các dự đoán là Low quality là đúng.
 - o Recall (0.15): Mô hình chỉ nhận diện đúng 15% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.16): Hiệu suất rất thấp, cho thấy mô hình gặp khó khăn trong việc phân loại lớp này, có thể do số lượng mẫu quá ít.
- Lớp Standard quality:
 - o Precision (0.84): 84% các dự đoán là Standard quality là chính xác.
 - o Recall (0.85): Nhận diện đúng 85% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.84): Hiệu suất cao, cho thấy mô hình hoạt động rất tốt đối với lớp chiếm ưu thế này.
- Lớp High quality:
 - o Precision (0.51): 51% các dự đoán là High quality là chính xác.
 - o Recall (0.52): Mô hình nhận diện được 52% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.51): Hiệu suất trung bình, nhưng còn cần cải thiện.
- Trung bình:
 - o Macro Average:
 - Precision (0.51): Hiệu suất trung bình giữa ba lớp, chỉ ra rằng mô hình bị ảnh hưởng bởi lớp Low quality có Precision thấp.
 - Recall (0.50): Mức độ nhận diện trung bình giữa các lớp.
 - F1-Score (0.51): Hiệu suất tổng quan chưa thực sự cao do mất cân bằng lớp.
 - o Weighted Average:
 - Precision (0.75): Hiệu suất trung bình có trọng số, phản ánh chính xác hơn nhờ lớp Standard quality chiếm ưu thế.
 - Recall (0.75): Tỷ lệ nhận diện chính xác trên toàn bộ tập dữ liệu.
 - F1-Score (0.75): Hiệu suất ổn định nhờ đóng góp lớn từ lớp chiếm đa số.



- Nhận xét: Với tỉ lệ tập train/test là 40/60:

- Precision: Lớp Standard quality đạt Precision cao (0.84), trong khi lớp Low quality (0.18) và High quality (0.51) thấp, do mất cân bằng dữ liệu.
- Recall: Lớp Standard quality có Recall cao (0.85), nhưng lớp Low quality (0.15) và High quality (0.52) còn hạn chế.
- F1-Score và Accuracy: Lớp Standard quality đạt F1-Score cao (0.84), giúp Accuracy tổng thể đạt 75%, nhưng hiệu suất ở các lớp nhỏ vẫn kém.
- Kết luận: Mô hình hoạt động tốt với lớp chiếm ưu thế (Standard quality), nhưng cần xử lý mất cân bằng để cải thiện các lớp Low quality và High quality.

Bộ dữ liệu 60/40 (Train/Test)

Train/Test Split: 60/40, Lần chia thứ 1

Classification Report:

	precision	recall	f1-score	support
Low quality	0.23	0.28	0.25	82
Standard quality	0.85	0.84	0.85	1504
High quality	0.56	0.57	0.56	374
accuracy			0.76	1960
macro avg	0.55	0.56	0.55	1960
weighted avg	0.77	0.76	0.77	1960

- Hiệu suất tổng thể: Mô hình đạt độ chính xác (Accuracy) là 76%, tức dự đoán đúng 76% số mẫu trong tập kiểm tra.
- Lớp Low quality:
 - o Precision (0.23): Chỉ 23% các dự đoán là Low quality là chính xác.
 - o Recall (0.28): Mô hình chỉ nhận diện được 28% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.25): Hiệu suất rất thấp, do mất cân bằng dữ liệu nghiêm trọng.
- Lớp Standard quality:
 - o Precision (0.85): Dự đoán lớp Standard quality chính xác 85%.
 - o Recall (0.84): Nhận diện được 84% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.85): Hiệu suất cao, mô hình hoạt động tốt với lớp chiếm ưu thế này.
- Lớp High quality:
 - o Precision (0.56): 56% các dự đoán là High quality là chính xác.
 - o Recall (0.57): Nhận diện được 57% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.56): Hiệu suất trung bình, cần cải thiện.
- Trung bình:
 - o Macro Average: Hiệu suất trung bình giữa ba lớp: Precision 0.55, Recall 0.56, F1-Score 0.55.
 - o Weighted Average: Hiệu suất trung bình có trọng số (dựa trên số mẫu mỗi lớp): Precision 0.77, Recall 0.76, F1-Score 0.77.



- Nhận xét: Với tỉ lệ tập train/test là 60/40:
 - Precision: Lớp Standard quality có Precision cao (0.85), trong khi lớp Low quality (0.23) và High quality (0.56) thấp, đặc biệt là lớp Low quality.
 - Recall: Lớp Standard quality đạt Recall tốt (0.84), nhưng lớp Low quality (0.28) và High quality (0.57) còn hạn chế.
 - F1-Score và Accuracy: F1-Score của lớp Standard quality đạt cao nhất (0.85), hỗ trợ Accuracy tổng thể đạt 76%, nhưng hiệu suất kém ở lớp Low quality kéo trung bình xuống.
 - Kết luận: Mô hình dự đoán tốt với lớp Standard quality, nhưng cần cải thiện hiệu suất ở các lớp Low quality và High quality để đảm bảo cân bằng.

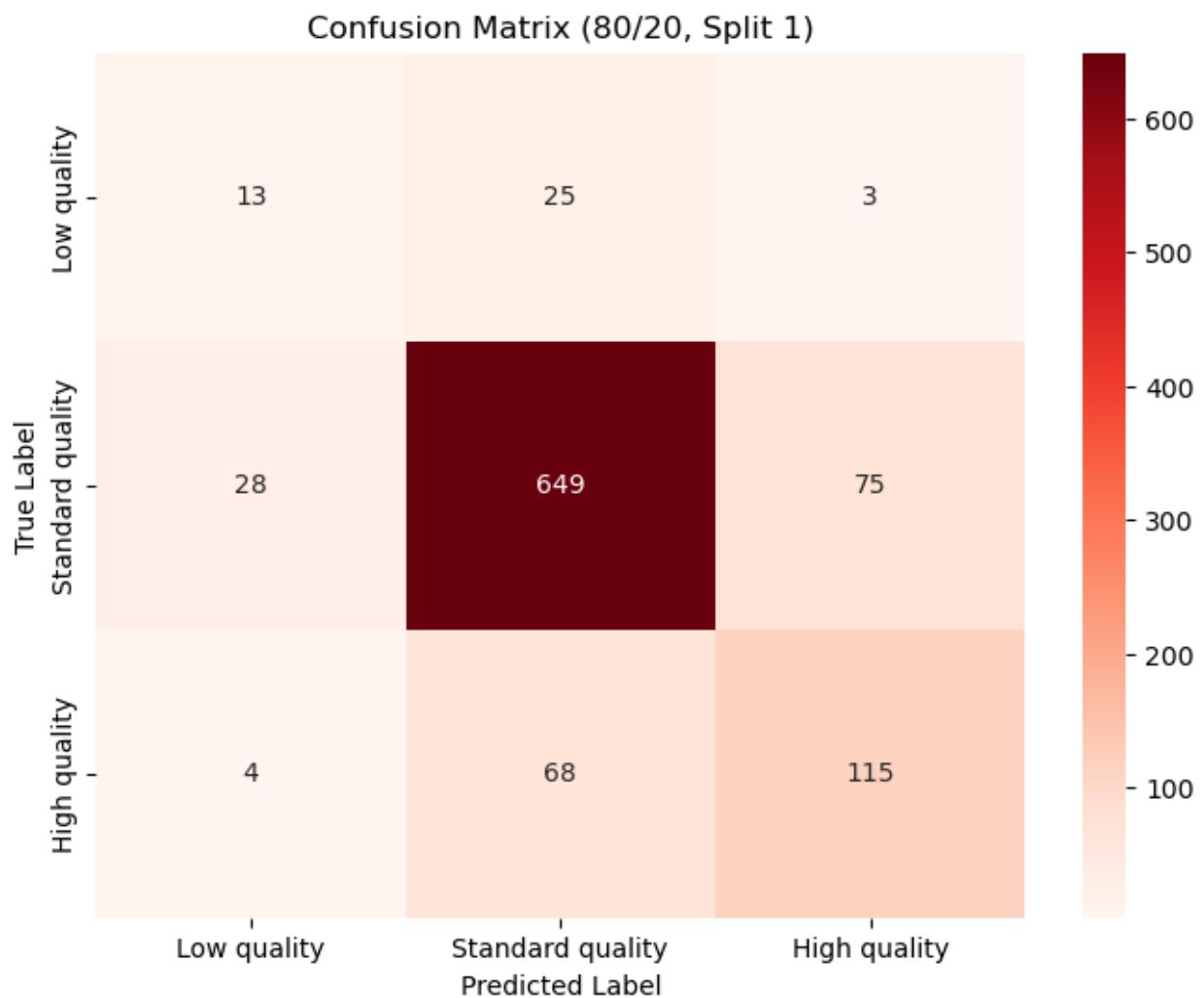
Bộ dữ liệu 80/20 (Train/Test)

Train/Test Split: 80/20, Lần chia thứ 1

Classification Report:

	precision	recall	f1-score	support
Low quality	0.29	0.32	0.30	41
Standard quality	0.87	0.86	0.87	752
High quality	0.60	0.61	0.61	187
accuracy			0.79	980
macro avg	0.59	0.60	0.59	980
weighted avg	0.80	0.79	0.79	980

- Hiệu suất tổng thể: Mô hình đạt độ chính xác (Accuracy) là 79%, tức dự đoán đúng 79% số mẫu trong tập kiểm tra.
- Lớp Low quality:
 - o Precision (0.29): Chỉ 29% các dự đoán là Low quality là chính xác.
 - o Recall (0.32): Mô hình nhận diện được 32% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.30): Hiệu suất thấp, do số lượng mẫu ít và mất cân bằng dữ liệu.
- Lớp Standard quality:
 - o Precision (0.87): Dự đoán lớp Standard quality chính xác 87%.
 - o Recall (0.86): Nhận diện được 86% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.87): Hiệu suất cao, cho thấy mô hình hoạt động tốt với lớp chiếm đa số này.
- Lớp High quality:
 - o Precision (0.60): 60% các dự đoán là High quality là chính xác.
 - o Recall (0.61): Nhận diện được 61% các mẫu thực sự thuộc lớp này.
 - o F1-Score (0.61): Hiệu suất trung bình, cần cải thiện.
- Trung bình:
 - o Macro Average: Hiệu suất trung bình giữa ba lớp: Precision 0.59, Recall 0.60, F1-Score 0.59, phản ánh sự chênh lệch hiệu suất giữa các lớp.
 - o Weighted Average: Hiệu suất trung bình có trọng số (dựa trên số mẫu mỗi lớp): Precision 0.80, Recall 0.79, F1-Score 0.79, được cải thiện nhờ lớp chiếm ưu thế.



- Nhận xét: Với tỉ lệ tập train/test là 80/20:

- Precision:
 - Precision của lớp Standard quality là 0.87, cao hơn so với lớp High quality (0.60) và lớp Low quality (0.29).
 - Điều này cho thấy mô hình dự đoán lớp chiếm ưu thế (Standard quality) tốt hơn hẳn, nhưng bị giảm hiệu suất với các lớp ít mẫu.
- Recall:
 - Recall của lớp Low quality là 0.32, thấp hơn nhiều so với lớp Standard quality (0.86) và lớp High quality (0.61).
 - Mô hình ưu tiên nhận diện mẫu thuộc lớp Standard quality, trong khi khả năng nhận diện lớp Low quality vẫn hạn chế đáng kể.
- F1-Score và Accuracy:
 - F1-Score của lớp Standard quality (0.87) cao hơn hẳn các lớp khác, đặc biệt là lớp Low quality (0.30).
 - Accuracy đạt 79% phản ánh mô hình vẫn đạt hiệu quả phân loại tốt nhưng cần cải thiện ở các lớp nhỏ.
- Kết luận: Mô hình đạt hiệu suất cao với lớp Standard quality, nhưng khả năng nhận diện và dự đoán các lớp Low quality và High quality còn yếu. Hiệu suất tổng thể tốt, nhưng cần xử lý mất cân bằng để cải thiện.

Bộ dữ liệu 90/10 (Train/Test)

Train/Test Split: 90/10, Lần chia thứ 1

Classification Report:

	precision	recall	f1-score	support
Low quality	0.29	0.25	0.27	20
Standard quality	0.85	0.88	0.87	376
High quality	0.61	0.54	0.57	94
accuracy			0.79	490
macro avg	0.58	0.56	0.57	490
weighted avg	0.78	0.79	0.79	490

- Hiệu suất tổng thể: : Mô hình đạt độ chính xác (Accuracy) là 79%, tức dự đoán đúng 79% số mẫu trong tập kiểm tra.
- Lớp Low quality:
 - o Precision (0.29): Dự đoán lớp "Low quality" chính xác chỉ 29%. Đây là một kết quả rất thấp, cho thấy mô hình gặp khó khăn trong việc phân loại chính xác các mẫu thuộc lớp này.
 - o Recall (0.25): Nhận diện đúng 25% các mẫu thực sự thuộc lớp "Low quality". Mô hình chưa thể nhận diện tốt lớp này, có thể do sự mất cân bằng dữ liệu.
 - o F1-Score (0.27): Hiệu suất cân bằng giữa Precision và Recall rất thấp, cho thấy mô hình không hoạt động tốt với lớp này.
- Lớp Standard quality:
 - o Precision (0.85): Dự đoán lớp Standard quality chính xác 85%. Đây là mức độ dự đoán rất cao, cho thấy mô hình xử lý tốt lớp này.
 - o Recall (0.88): Nhận diện đúng 88% các mẫu thực sự thuộc lớp "Standard quality". Mô hình khá hiệu quả trong việc nhận diện lớp này.
 - o F1-Score (0.87): Hiệu suất giữa Precision và Recall rất cao, cho thấy mô hình hoạt động rất tốt với lớp này.
- Lớp High quality:
 - o Precision (0.61): Dự đoán lớp High quality chính xác 61%. Mặc dù không thấp, nhưng còn có thể cải thiện.
 - o Recall (0.54): Nhận diện đúng 54% các mẫu thực sự thuộc lớp "High quality". Khả năng nhận diện lớp này chưa thật sự mạnh mẽ.
 - o F1-Score (0.57): Hiệu suất cân bằng giữa Precision và Recall chỉ ở mức trung bình, cho thấy mô hình vẫn cần cải thiện để xử lý tốt lớp này.
- Trung bình:
 - o Macro Average: Precision: 0.58, Recall: 0.56, F1-Score: 0.57. Hiệu suất trung bình giữa các lớp thấp, điều này phần lớn do sự chênh lệch trong số lượng mẫu giữa các lớp.
 - o Weighted Average: Precision: 0.78, Recall: 0.79, F1-Score: 0.79. Hiệu suất trung bình có trọng số cao hơn, nhờ vào sự đóng góp lớn từ lớp Standard quality".

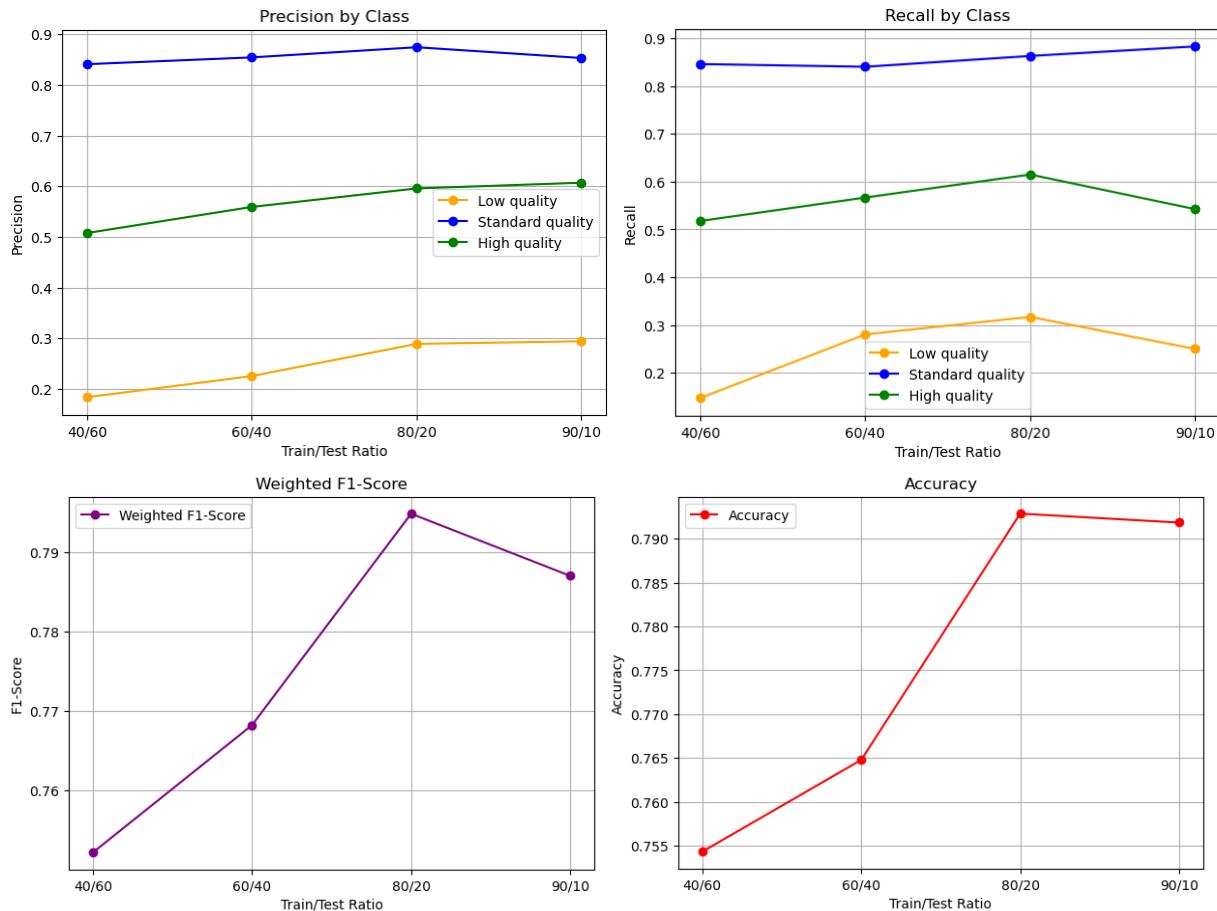


- Nhận xét: Với tỉ lệ tập train/test là 90/10:

- Precision của lớp Standard quality là 0.85 và của lớp Low quality là 0.29. Tỉ lệ này có sự thiên lệch nhiều hơn so với các lần chia trước, với lớp Standard quality đạt kết quả rất cao nhưng lớp Low quality rất thấp.
- Recall của lớp Low quality là 0.25 thấp hơn nhiều so với lớp Standard quality (0.88), cho thấy mô hình ưu tiên nhận diện mẫu thuộc lớp Standard quality hơn lớp Low quality. Khả năng nhận diện mẫu thuộc lớp Standard quality rất tốt, trong khi khả năng nhận diện mẫu của lớp Low quality giảm mạnh.
- F1-Score và Accuracy của lớp Standard quality cao, nhưng hiệu suất tổng thể vẫn thấp nhất trong các lần chia trước, cho thấy mô hình có hiệu quả phân loại khá tốt, nhưng cần cải thiện khả năng nhận diện các lớp ít phổ biến như Low quality.

Insight

Với dữ liệu quan sát ở trên, nhóm chọn ra các thông tin là precision, recall và accuracy là những thông tin có sự thay đổi đáng quan sát



Final Metrics Summary:

Train/Test Ratio: 40/60

Precision: Low=0.1837, Standard=0.8413, High=0.5079

Recall: Low=0.1475, Standard=0.8461, High=0.5178

Weighted F1-Score: 0.7522

Accuracy: 0.7543

Train/Test Ratio: 60/40

Precision: Low=0.2255, Standard=0.8546, High=0.5594

Recall: Low=0.2805, Standard=0.8404, High=0.5668

Weighted F1-Score: 0.7682

Accuracy: 0.7648

Train/Test Ratio: 80/20

Precision: Low=0.2889, Standard=0.8747, High=0.5959

Recall: Low=0.3171, Standard=0.8630, High=0.6150

Weighted F1-Score: 0.7948

Accuracy: 0.7929

Train/Test Ratio: 90/10

Precision: Low=0.2941, Standard=0.8535, High=0.6071

Recall: Low=0.2500, Standard=0.8830, High=0.5426

Weighted F1-Score: 0.7870

Accuracy: 0.7918

- Với lớp Low quality, khi tỉ lệ tập test giảm từ 0.6 xuống 0.1, precision và recall của lớp này tăng dần. Tuy nhiên, mức precision và recall của lớp Low quality vẫn thấp, đặc biệt ở các tỉ lệ test lớn, điều này cho thấy mô hình gặp khó khăn trong việc phân loại chính xác các mẫu thuộc lớp này. Mặc dù có sự cải thiện khi giảm tỉ lệ tập test, nhưng hiệu suất phân loại lớp Low quality vẫn rất hạn chế.
- Với lớp Standard quality, precision và recall của lớp này luôn duy trì ở mức cao và ổn định. Khi tỉ lệ test giảm từ 0.6 xuống 0.1, precision và recall đều ở mức cao nhất cho lớp này, cho thấy mô hình phân loại chính xác lớp Standard quality rất tốt, đặc biệt khi tỉ lệ test thấp hơn.
- Với lớp High quality, precision và recall của lớp này cũng tăng nhẹ khi tỉ lệ test giảm, tuy nhiên vẫn không bằng lớp Standard quality. Mặc dù mức precision và recall có sự cải thiện nhẹ, mô hình vẫn gặp khó khăn trong việc phân loại chính xác các mẫu thuộc lớp này, đặc biệt khi tỉ lệ test giảm xuống thấp.
- Hiệu suất tổng thể: Accuracy và weighted F1-score tăng dần khi tỉ lệ test giảm từ 0.6 xuống 0.1, nhưng vẫn có sự chênh lệch đáng kể giữa các lớp. Mặc dù mô hình duy trì được hiệu suất ổn định, lớp Low quality vẫn là lớp khó phân loại, trong khi lớp Standard quality và High quality có độ chính xác cao hơn.
- Nhìn chung, mô hình hoạt động tốt nhất với lớp Standard quality, nhưng gặp khó khăn khi phân loại lớp Low quality, đặc biệt khi tỉ lệ test lớn. Sự thiên lệch trong dữ liệu và sự chênh lệch giữa các lớp có thể là nguyên nhân khiến mô hình phân loại lớp Low quality kém hiệu quả hơn.

2.2.5 Phân tích độ sâu và độ chính xác

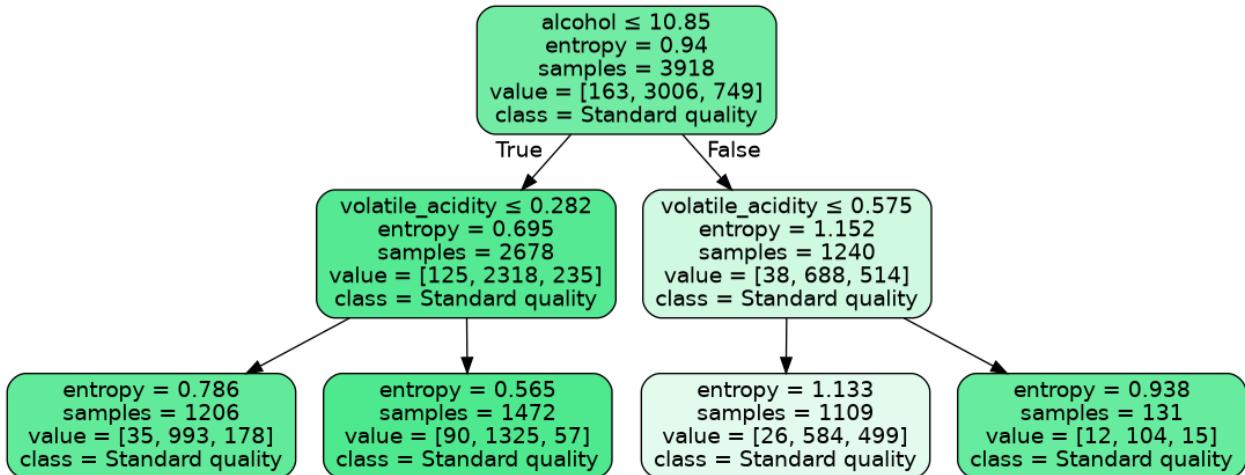
Max depth = None (Không có độ sâu giới hạn)

File ảnh này khá phức tạp và khó để thể hiện trong báo cáo. Nhóm em xin phép chỉ đưa đường dẫn a.

Đường dẫn: Source\Wine\Image4Depths\decision_tree_max_depth_None.png

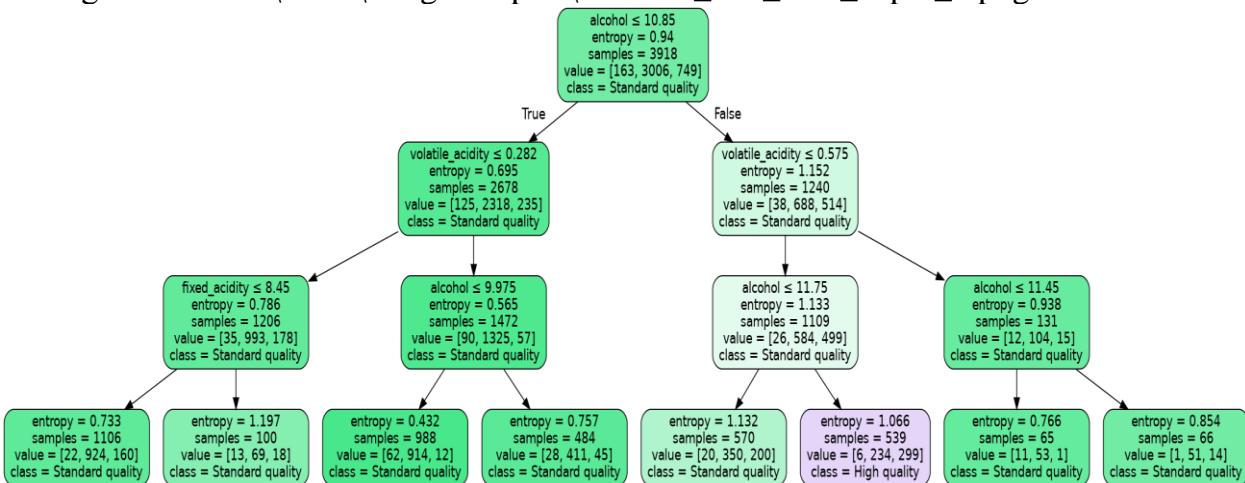
Max depth = 2

Đường dẫn: Source\Wine\Image4Depths\decision_tree_max_depth_2.png



Max depth = 3

Đường dẫn: Source\Wine\Image4Depths\decision_tree_max_depth_3.png



Max depth = 4

Đường dẫn: Source\Wine\Image4Depths\decision_tree_max_depth_4.png

Max depth = 5

Đường dẫn: Source\Wine\Image4Depths\decision_tree_max_depth_5.png

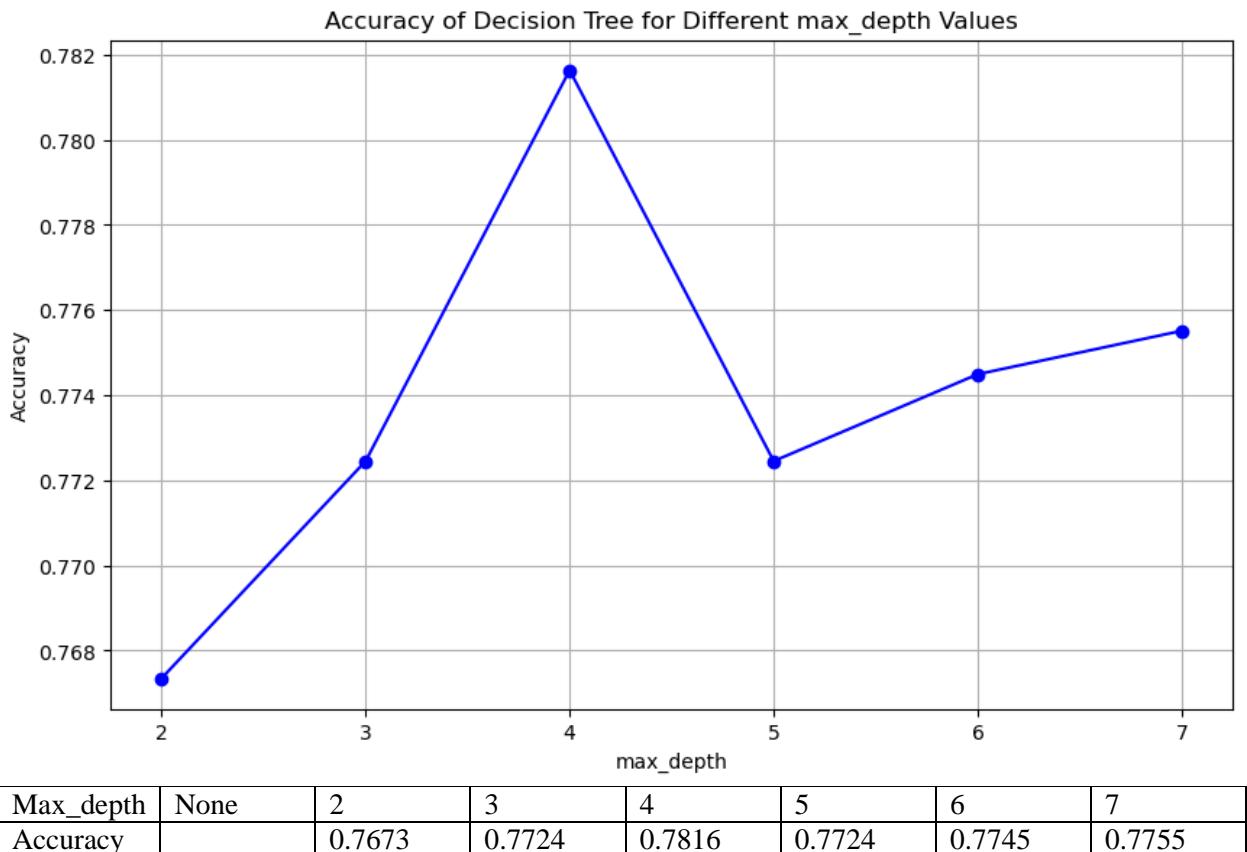
Max depth = 6

Đường dẫn: Source\Wine\Image4Depths\decision_tree_max_depth_7.png

Max depth = 7

Đường dẫn: Source\Wine\Image4Depths\decision_tree_max_depth_7.png

So sánh các độ sâu



Với độ sâu là None, nhóm em không chạy được thuật toán vì bị lỗi : Not supported between of “int” and “NoneType”. Dù đã thử nhiều cách nhưng vẫn không thực thi được, chúng em đoán có lẽ là bị lỗi dữ liệu đầu vào và xin phép thầy được bỏ qua phần số liệu cho None.

Insight:

- Hiệu suất tổng thể: Accuracy dao động từ 0.7673 đến 0.7816, cho thấy mô hình đạt hiệu suất phân loại ở mức khá ổn định với các giá trị khác nhau của Max Depth.
- Ảnh hưởng của Max Depth:
 - o Max_depth = 2: Accuracy thấp nhất (0.7673), có thể do cây quá nông, chưa đủ khả năng phân biệt tốt các đặc trưng của dữ liệu.
 - o Max_depth = 4: Accuracy đạt cao nhất (0.7816), cho thấy đây là giá trị tối ưu, giúp mô hình cân bằng giữa việc tổng quát hóa và phân loại chính xác trên tập kiểm tra.
 - o Max_depth = 5-7: Accuracy giảm nhẹ sau độ sâu 4, dao động quanh 0.7724-0.7755, điều này có thể do mô hình bắt đầu overfitting khi cây học quá chi tiết dữ liệu huấn luyện.
- Xu hướng:
 - o Tăng Max Depth từ 2 lên 4 giúp cải thiện hiệu suất rõ rệt, nhưng sau đó Accuracy không tăng thêm mà giảm nhẹ.
 - o Điều này cho thấy việc giới hạn độ sâu tại 4 là cần thiết để tránh overfitting và duy trì hiệu suất tốt nhất.

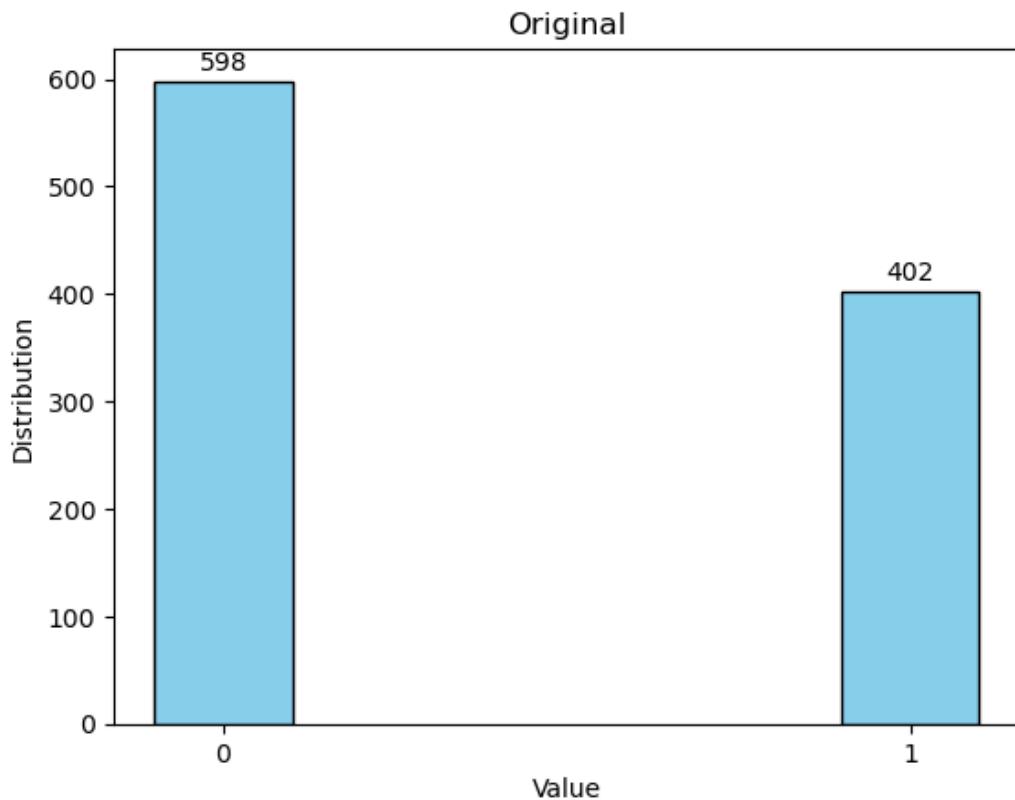
- Kết luận:

- Max_depth = 4 là giá trị tối ưu cho dataset này, đạt Accuracy cao nhất (0.7816) và đảm bảo cân bằng giữa độ phức tạp và khả năng tổng quát của mô hình.
- Tăng Max Depth quá mức không mang lại lợi ích rõ rệt và có thể dẫn đến overfitting. Do đó, cần ưu tiên giới hạn độ sâu để đạt hiệu suất tốt nhất..

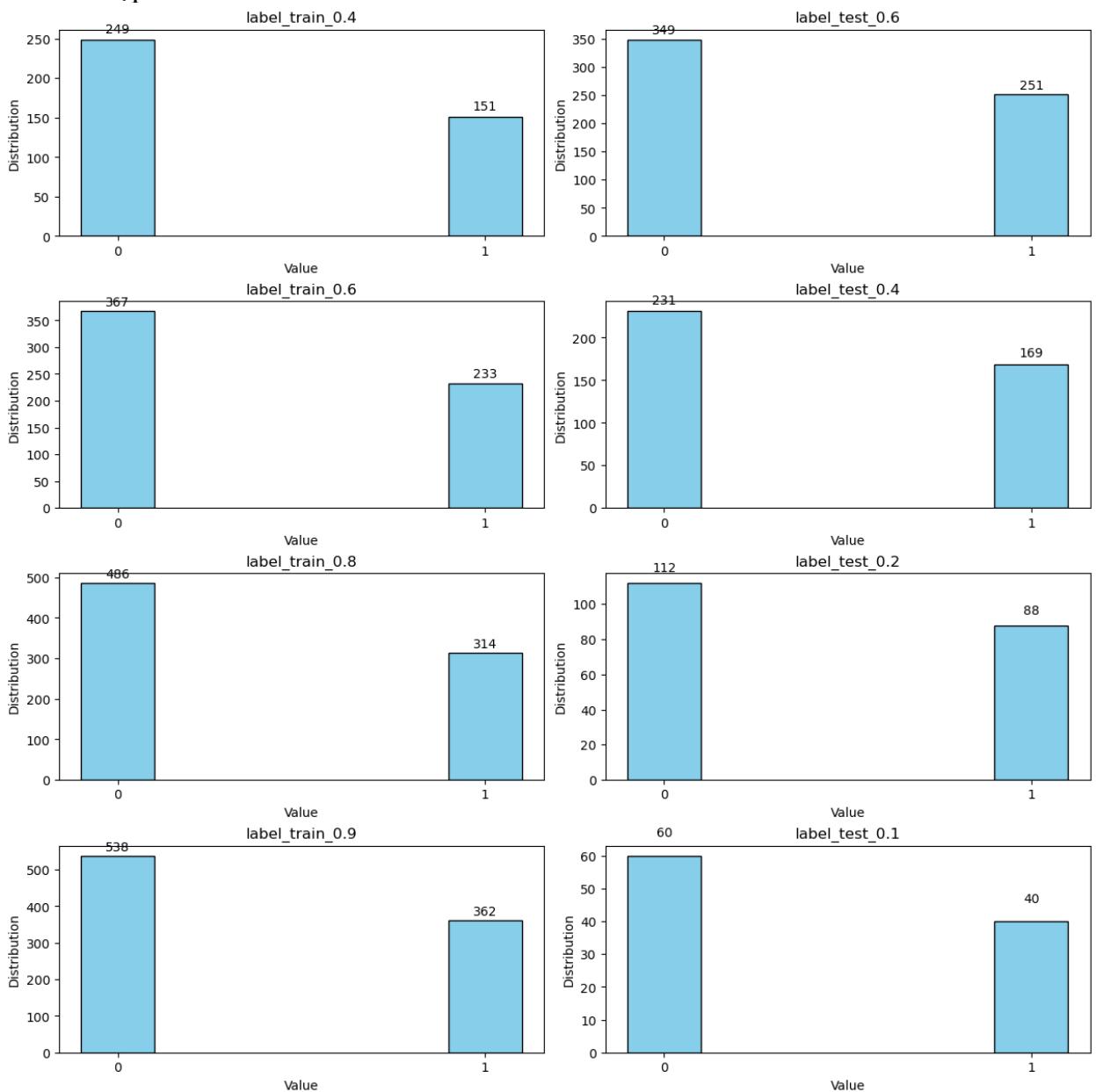
2.3 Additional Dataset - Car Purchase Dataset

2.3.1 Chuẩn bị dữ liệu

- Bộ dữ liệu sử dụng trong dự án này là Car Purchase Dataset, được lấy từ website “kaggle” với tên trang web là “Cars - Purchase Decision Dataset” của tác giả “Gabriel Santello”
- Link trang web: <https://www.kaggle.com/datasets/gabrielsantello/cars-purchase-decision-dataset>
- Đường dẫn file csv trong thư mục: Source\Addition\car_data.csv
- Bộ dữ liệu chứa thông tin về khách hàng và quyết định mua xe với số đặc trưng (Features) là 3, bao gồm:
 - o Giới tính (Gender).
 - o Tuổi (Age).
 - o Thu nhập (AnnualSalary).
- Số mẫu (Samples): 1000
- Đầu ra (Targets): Quyết định mua xe của khách hàng, được biểu diễn dưới dạng nhị phân:
 - o 0: Không mua hàng (No buy).
 - o 1: Có mua hàng (Buy).
- Dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo các tỷ lệ khác nhau (40%, 60%, 80%, và 90%) để đánh giá hiệu suất của mô hình dưới các kích thước tập huấn luyện khác nhau. Dùng `train_test_split` để tạo tập huấn luyện và kiểm tra, đảm bảo phân phối lớp giữa các tập giống nhau dữ liệu ban đầu.
- Kết quả được lưu với tên: `feature_train`, `feature_test`, `label_train`, `label_test`.
- Trực quan hóa các tập dữ liệu như sau
- o Tập ban đầu



○ Tập sau khi chia



Nhận xét

- Phân phối gốc (Original Dataset): Tỷ lệ các lớp
 - Lớp 1 (Buy): 402 mẫu – 40.2 %
 - Lớp 0 (No Buy): 598 mẫu – 59.8 %
- Tập dữ liệu ban đầu không cân bằng nhưng chênh lệch không lớn.
- Phân phối các tập con: Tỷ lệ các lớp trong tập huấn luyện và kiểm tra
 - Ở tất cả các tỷ lệ phân chia (40/60, 60/40, 80/20, 90/10), phân phối lớp trong cả tập huấn luyện và tập kiểm tra gần như giữ nguyên so với tập dữ liệu gốc

2.3.2 Cài đặt

- Việc cài đặt tương tự 2 dataset ở trên: dùng DecisionTreeClassifier để xây dựng cây và graphviz để trực quan hóa.



2.3.3 Trực quan hóa cây quyết định

Vì lí do ảnh khó có thể biểu diễn trong báo cáo nhóm xin phép chỉ đưa đường dẫn ảnh ạ.

Bộ dữ liệu 40/60 (Train/Test)

Đường dẫn: Source\Addition\Image4Datasets\data_40_60.png

Bộ dữ liệu 60/40 (Train/ Test)

Đường dẫn: Source\Addition\Image4Datasets\data_60_40.png

Bộ dữ liệu 80/20 (Train/Test)

Đường dẫn: Source\Addition\Image4Datasets\data_80_20.png

Bộ dữ liệu 90/10 (Train/Test)

Đường dẫn: Source\Addition\Image4Datasets\data_90_10.png

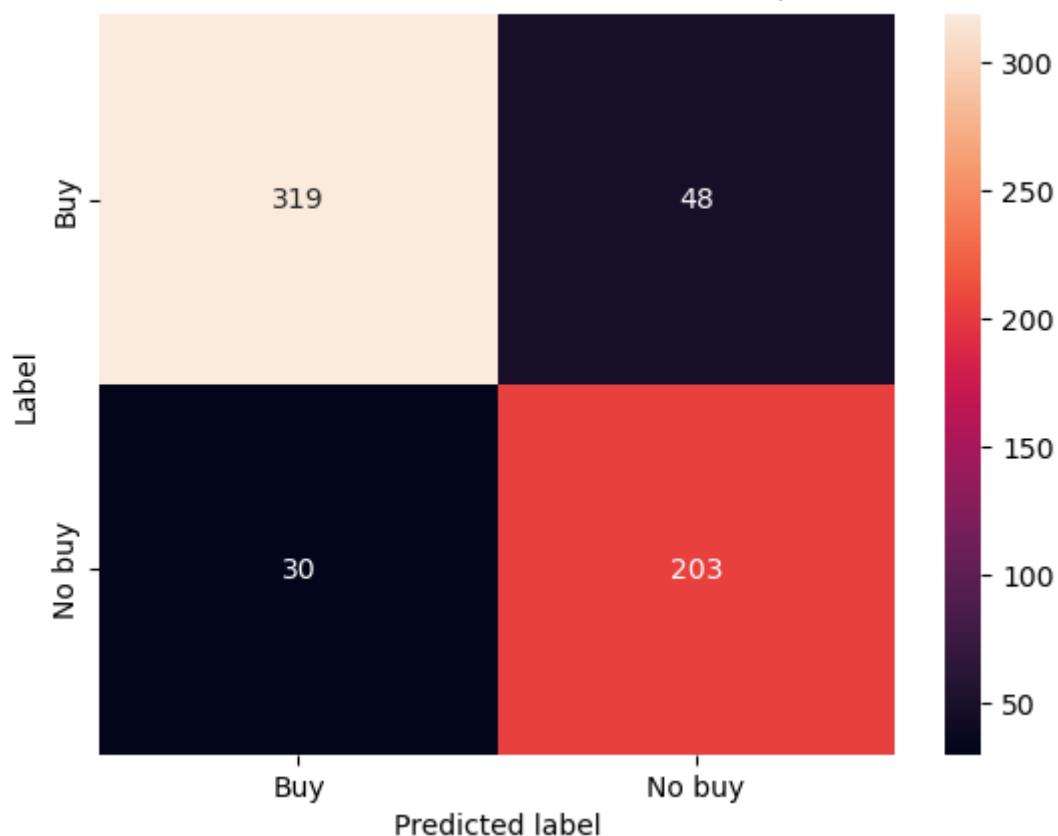
2.3.4 Đánh giá cây quyết định

Bộ dữ liệu 40/60 (Train/Test)

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.87	0.89	367
1	0.81	0.87	0.84	233
accuracy			0.87	600
macro avg	0.86	0.87	0.86	600
weighted avg	0.87	0.87	0.87	600

Decision Tree Classifier confusion matrix - 40/60 dataset



- Nhận xét: Với tỉ lệ tập train/test là 40/60:

o Precision:

- Precision của lớp 0 là 0.91, nghĩa là 91% các dự đoán thuộc lớp 0 là chính xác.
- Precision của lớp 1 là 0.81, thấp hơn lớp 0, cho thấy mô hình dự đoán nhầm một số mẫu không thuộc lớp 1 thành lớp này.

o Recall:

- Recall của lớp 0 là 0.87, nghĩa là mô hình nhận diện đúng 87% các mẫu thực sự thuộc lớp này.
- Recall của lớp 1 cũng đạt 0.87, cho thấy khả năng nhận diện mẫu giữa hai lớp là cân bằng.

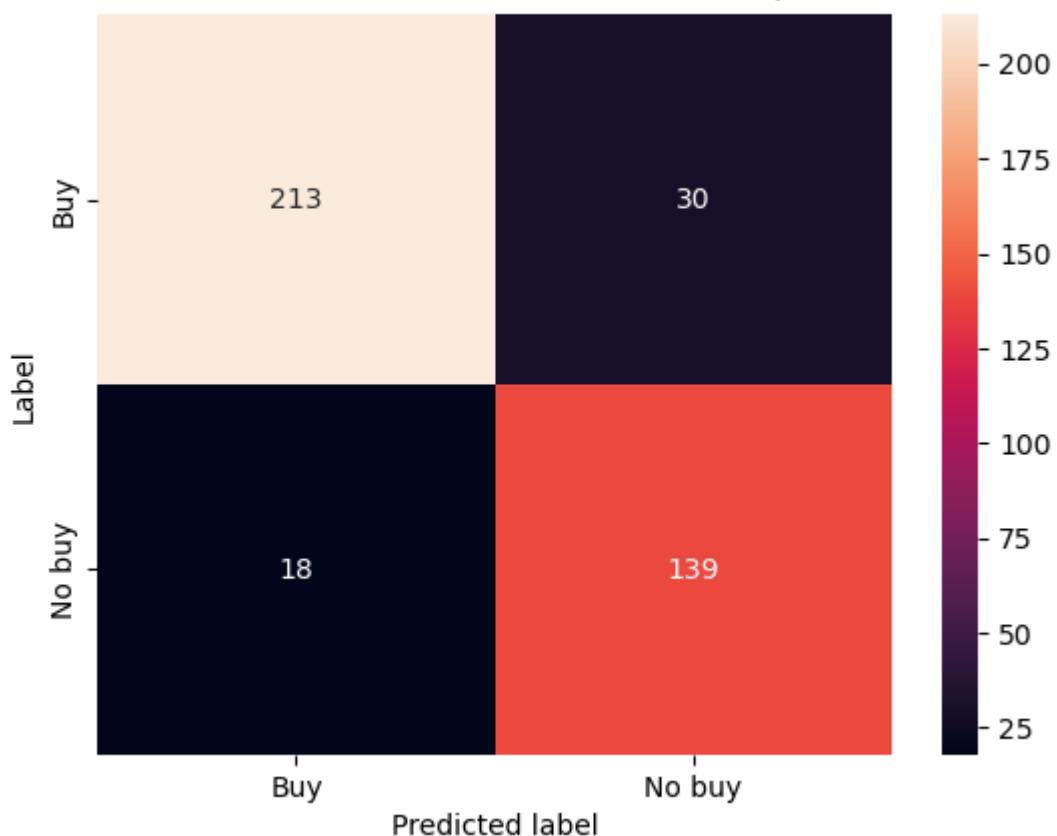
- F1-Score và Accuracy:
 - F1-Score của lớp 0 là 0.89 và của lớp 1 là 0.84, phản ánh hiệu suất tốt trên cả hai lớp, đặc biệt ở lớp 0.
 - Accuracy đạt 87%, tức mô hình dự đoán đúng 87% tổng số mẫu.
- Trung bình:
 - Macro avg: Precision, Recall, và F1-Score trung bình giữa hai lớp đều khoảng 0.86-0.87, cho thấy hiệu suất tương đối cân đối giữa các lớp.
 - Weighted avg: Precision, Recall, và F1-Score cũng đạt 0.87, nhờ tỷ lệ mẫu giữa hai lớp không quá chênh lệch.

Bộ dữ liệu 60/40 (Train/Test)

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.88	0.90	243
1	0.82	0.89	0.85	157
accuracy			0.88	400
macro avg	0.87	0.88	0.88	400
weighted avg	0.88	0.88	0.88	400

Decision Tree Classifier confusion matrix - 60/40 dataset



- Nhận xét: Với tỉ lệ tập train/test là 60/40:

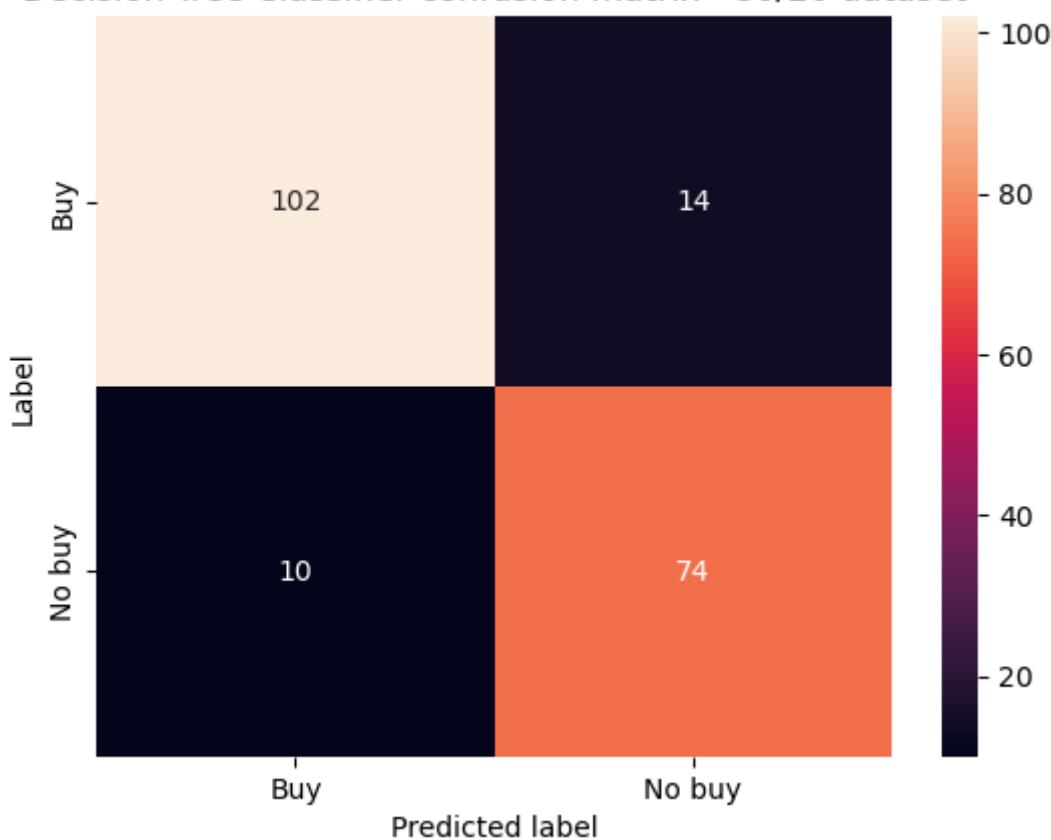
- Precision:
 - Precision của lớp 0 là 0.92, nghĩa là 92% các dự đoán thuộc lớp 0 là chính xác.
 - Precision của lớp 1 là 0.82, thấp hơn lớp 0, cho thấy mô hình có xu hướng dự đoán nhầm một số mẫu không thuộc lớp 1 thành lớp này.
- Recall:
 - Recall của lớp 0 là 0.88, tức mô hình nhận diện đúng 88% các mẫu thực sự thuộc lớp này.
 - Recall của lớp 1 là 0.89, cao hơn so với lớp 0, cho thấy mô hình nhận diện lớp 1 khá tốt.

- F1-Score và Accuracy:
 - F1-Score của lớp 0 là 0.90 và của lớp 1 là 0.85, phản ánh hiệu suất tổng quan tốt trên cả hai lớp.
 - Accuracy đạt 88%, chỉ ra rằng mô hình dự đoán đúng 88% tổng số mẫu.
- Trung bình:
 - Macro avg: Precision, Recall, và F1-Score trung bình giữa hai lớp đều khoảng 0.87-0.88, cho thấy hiệu suất cân đối giữa các lớp.
 - Weighted avg: Các chỉ số cũng đạt 0.88, nhờ tỷ lệ mẫu giữa hai lớp không quá chênh lệch.

Bộ dữ liệu 80/20 (Train/Test)

Classification Report:					
	precision	recall	f1-score	support	
0	0.91	0.88	0.89	116	
1	0.84	0.88	0.86	84	
accuracy			0.88	200	
macro avg	0.88	0.88	0.88	200	
weighted avg	0.88	0.88	0.88	200	

Decision Tree Classifier confusion matrix - 80/20 dataset



- Nhận xét: Với tỉ lệ tập train/test là 80/20:

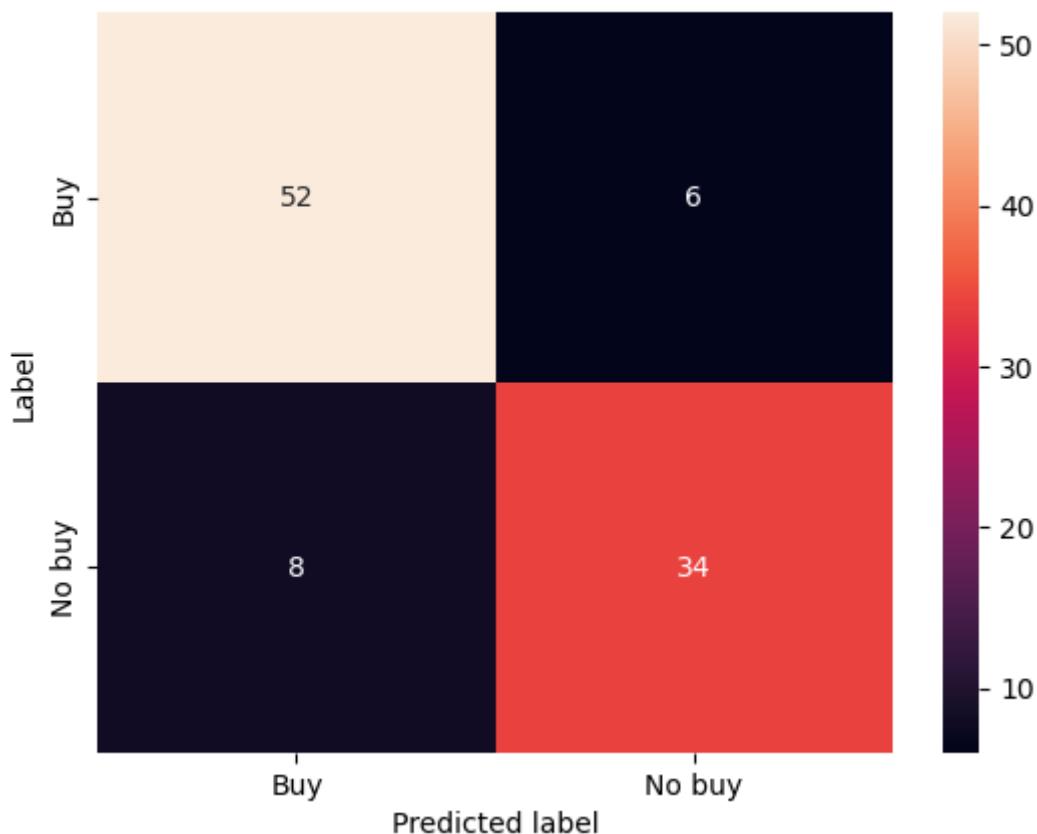
- o Precision: Precision của lớp 0 là 0.91 và của lớp 1 là 0.84, nghĩa là mô hình dự đoán chính xác khoảng 84-91% các mẫu thuộc cả hai lớp. Precision cao ở cả hai lớp, nhưng lớp 1 thấp hơn, cho thấy mô hình có xu hướng dự đoán nhầm một số mẫu không thuộc lớp 1.
- o Recall: Recall của lớp 0 là 0.88, tương đương với lớp 1 (0.88), cho thấy khả năng nhận diện mẫu thực sự thuộc hai lớp là cân bằng và tốt.
- o F1-Score và Accuracy:

- F1-Score của cả hai lớp lần lượt là 0.89 (lớp 0) và 0.86 (lớp 1), phản ánh hiệu suất tổng quan cao và ổn định.
- Accuracy đạt 88%, cho thấy mô hình hoạt động hiệu quả trong việc phân loại.
- Trung bình:
 - Macro avg: Precision, Recall, và F1-Score trung bình giữa hai lớp đều đạt 0.88, cho thấy hiệu suất cân đối giữa các lớp.
 - Weighted avg: Các chỉ số cũng đạt 0.88, nhờ phân phối số lượng mẫu giữa hai lớp không quá chênh lệch.

Bộ dữ liệu 90/10 (Train/Test)

Classification Report:					
	precision	recall	f1-score	support	
0	0.87	0.90	0.88	58	
1	0.85	0.81	0.83	42	
accuracy			0.86	100	
macro avg	0.86	0.85	0.86	100	
weighted avg	0.86	0.86	0.86	100	

Decision Tree Classifier confusion matrix - 90/10 dataset



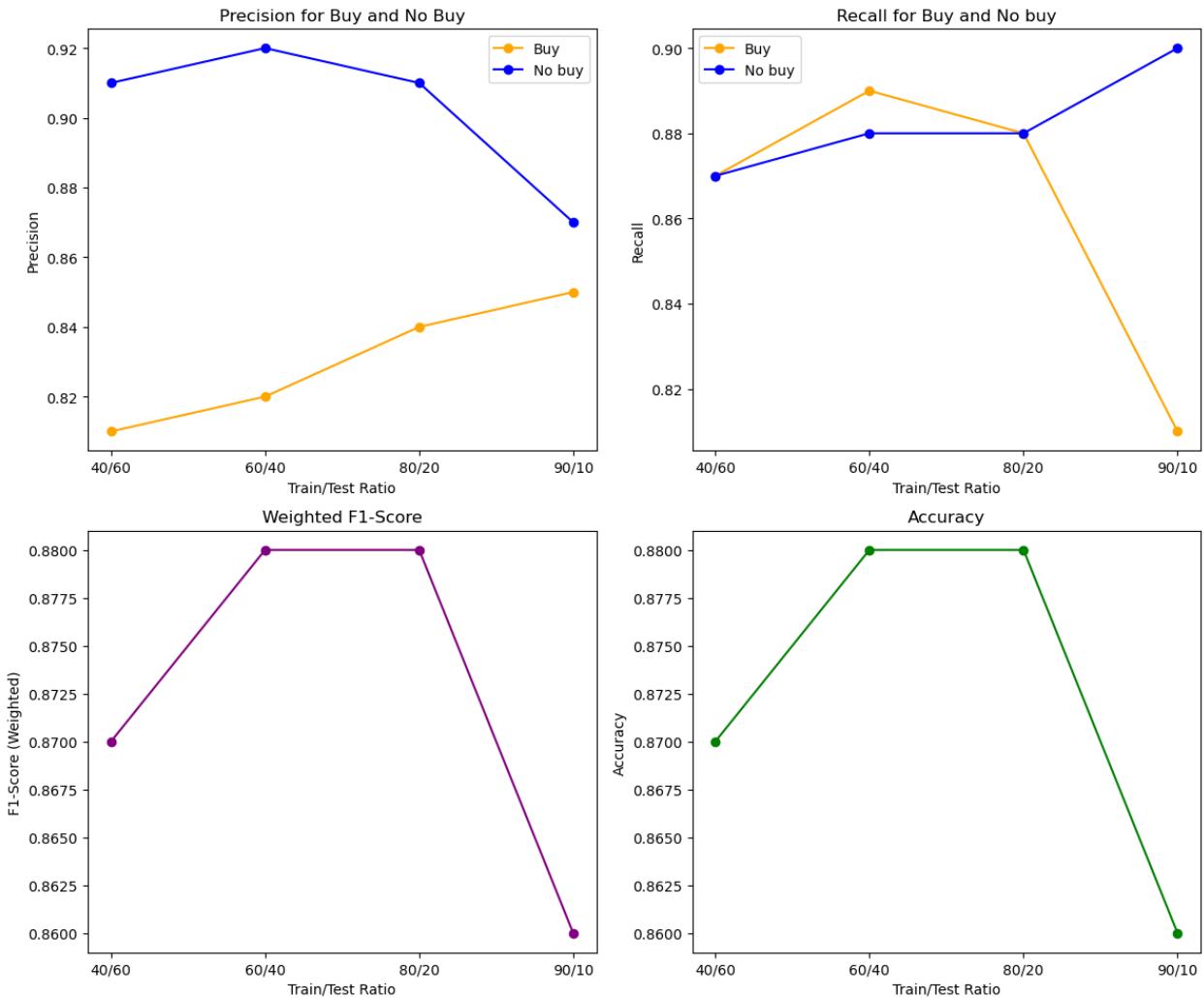
- Nhận xét: Với tỉ lệ tập train/test là 90/10:

- Precision: Precision của lớp 0 là 0.87 và của lớp 1 là 0.85, nghĩa là mô hình dự đoán chính xác khoảng 85-87% các mẫu thuộc cả hai lớp. Tỉ lệ này khá cao, cho thấy mô hình ít dự đoán nhầm giữa hai lớp.
- Recall: Recall của lớp 0 là 0.90, cao hơn so với lớp 1 (0.81). Tuy nhiên 2 tỉ lệ này không chênh lệch lớn và có kết quả tốt.
- F1-Score và Accuracy:
 - F1-Score của cả hai lớp lần lượt là 0.88 và 0.83, cho thấy hiệu suất tổng quan tốt.
 - Accuracy đạt 86%, phản ánh mô hình hoạt động hiệu quả trong việc phân loại.

- Trung bình:
 - Macro avg: Precision, Recall, và F1-Score trung bình giữa hai lớp đều khoảng 0.85-0.86, cho thấy hiệu suất cân đối giữa các lớp.
 - Weighted avg: Các chỉ số tương tự, nhờ số lượng mẫu giữa hai lớp không quá chênh lệch.
- Kết luận: Mô hình đạt hiệu suất phân loại tốt với độ chính xác khá cao (86% và 81%)

Insight

Với dữ liệu quan sát ở trên, nhóm chọn ra các thông tin là precision, recall và accuracy là những thông tin có sự thay đổi đáng quan sát



- Với lớp "No Buy" (0):
 - Precision duy trì ở mức cao (0.87-0.92) trong tất cả các tỷ lệ, cho thấy mô hình ít dự đoán nhầm các mẫu thuộc lớp khác thành lớp 0.
 - Recall dao động trong khoảng 0.87-0.90, nghĩa là mô hình nhận diện tương đối tốt các mẫu thuộc lớp này, đặc biệt ở tỷ lệ 90/10 (Recall cao nhất đạt 0.90).
 - F1-Score của lớp 0 ổn định (0.88-0.90) qua các tỷ lệ, cho thấy hiệu suất cân đối giữa Precision và Recall.
- Với lớp "Buy" (1):
 - Precision dao động từ 0.81 đến 0.85, với tỷ lệ 80/20 có giá trị cao nhất (0.85). Tỷ lệ 90/10 có Precision thấp hơn (0.85), cho thấy mô hình dễ nhầm các mẫu không thuộc lớp này khi tập test quá nhỏ.
 - Recall đạt mức cao nhất (0.89) ở tỷ lệ 60/40 và 80/20, nhưng giảm xuống 0.81 ở tỷ lệ 90/10, cho thấy khả năng nhận diện lớp 1 giảm khi dữ liệu kiểm tra quá ít.
 - F1-Score dao động từ 0.83-0.86, tốt nhất ở tỷ lệ 80/20, phản ánh sự cân bằng giữa Precision và Recall.

- Hiệu suất tổng thể: Accuracy ổn định ở mức cao (86%-88%) qua các tỷ lệ, với giá trị cao nhất tại tỷ lệ 60/40 và 80/20 (88%).
- Macro avg (trung bình giữa các lớp) và Weighted avg (trung bình có trọng số) đều tương đối ổn định (0.86-0.88).
- Nhận xét chi tiết:
 - Hiệu suất theo tỷ lệ train/test:
 - 40/60: Hiệu suất tốt, nhưng Precision của lớp 1 thấp hơn (0.81) so với các tỷ lệ khác, phản ánh khả năng nhận diện lớp Buy còn hạn chế.
 - 60/40 và 80/20: Đây là hai tỷ lệ có hiệu suất cao nhất (Accuracy và F1-Score đạt 88%). Mô hình đạt hiệu suất cân bằng giữa Precision và Recall của cả hai lớp.
 - 90/10: Hiệu suất giảm nhẹ (Accuracy 86%) do Precision và Recall của lớp 1 thấp hơn, có thể do tập test nhỏ dẫn đến sự thiên lèch.
 - Xu hướng chung:
 - Lớp No Buy (0) được dự đoán tốt hơn so với lớp Buy (1) trong tất cả các tỷ lệ, với Precision và F1-Score cao hơn.
 - Khi tỷ lệ test nhỏ hơn 0.2, mô hình gặp khó khăn trong việc nhận diện lớp Buy, thể hiện qua Precision và Recall của lớp này giảm.
 - Kết luận: Tỷ lệ train/test 60/40 và 80/20 mang lại hiệu suất tốt nhất, với độ chính xác (88%) và F1-Score cân bằng giữa hai lớp.

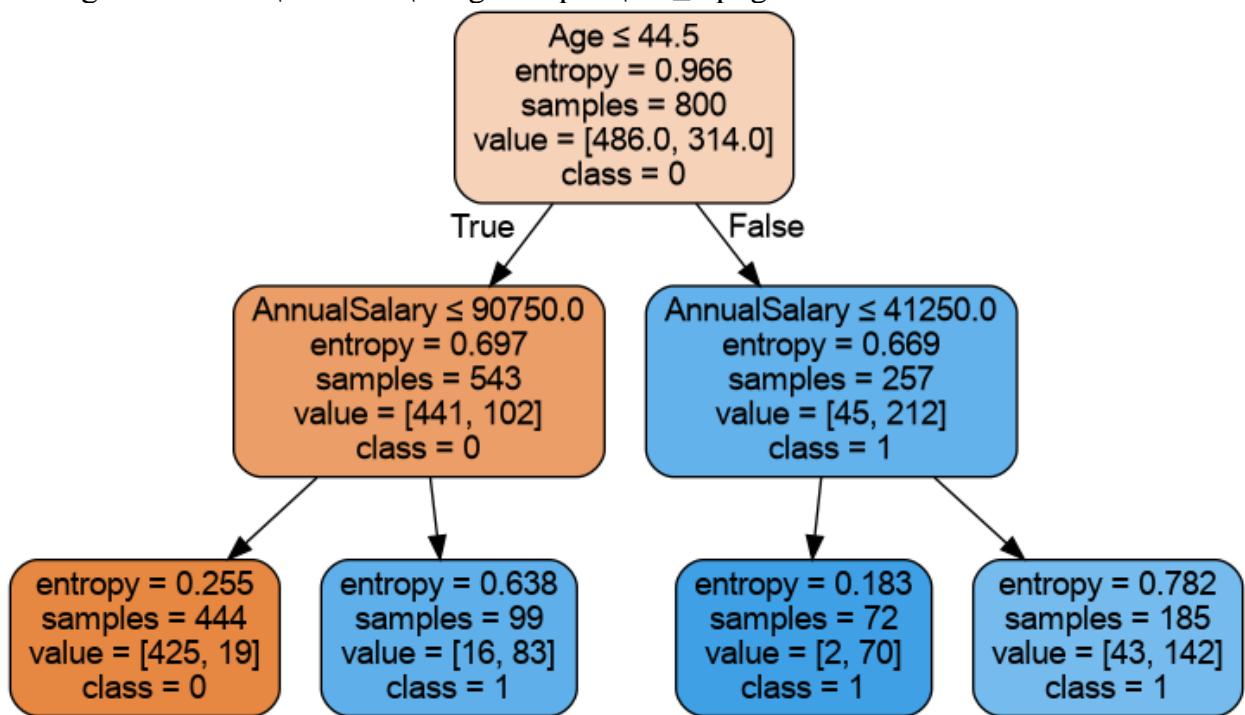
2.3.5 Phân tích độ sâu và độ chính xác

Max depth = None (Không có độ sâu giới hạn)

Đường dẫn: Source\Addition\Image7Depths\DT_Nonе.png

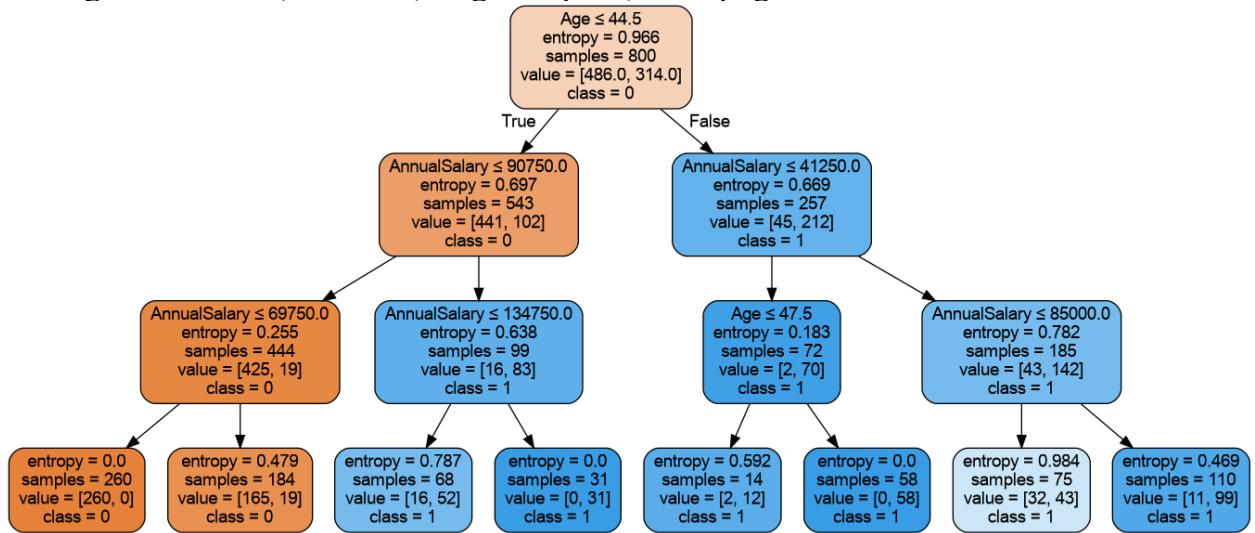
Max depth = 2

Đường dẫn: Source\Addition\Image7Depths\DT_2.png



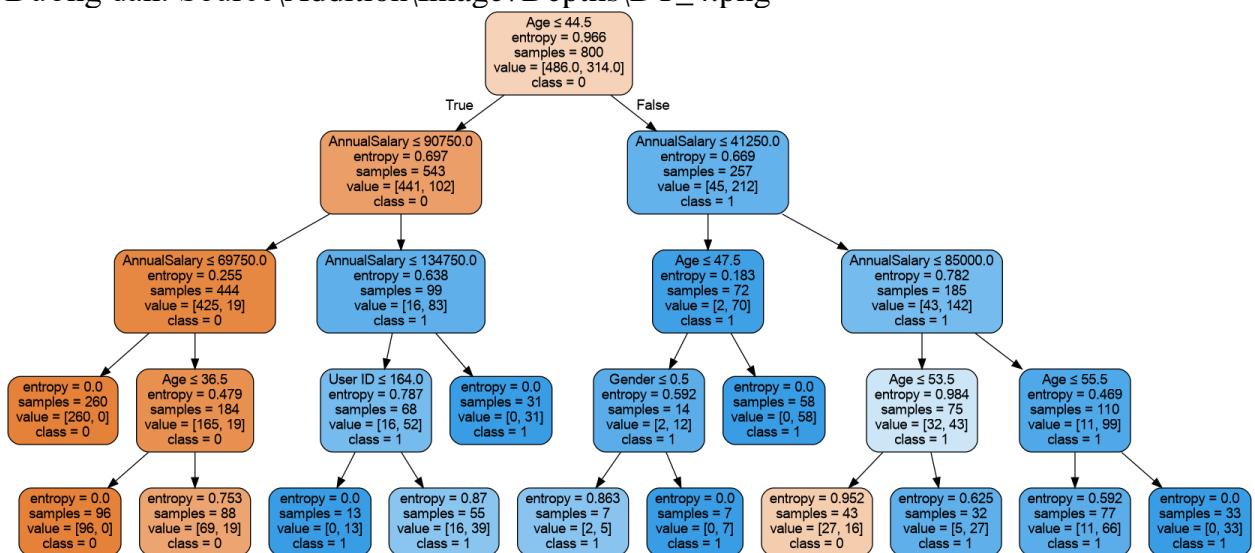
Max depth = 3

Đường dẫn: Source\Addition\Image7Depths\DT_3.png



Max depth = 4

Đường dẫn: Source\Addition\Image7Depths\DT_4.png



Max depth = 5

Đường dẫn: Source\Addition\Image7Depths\DT_5.png

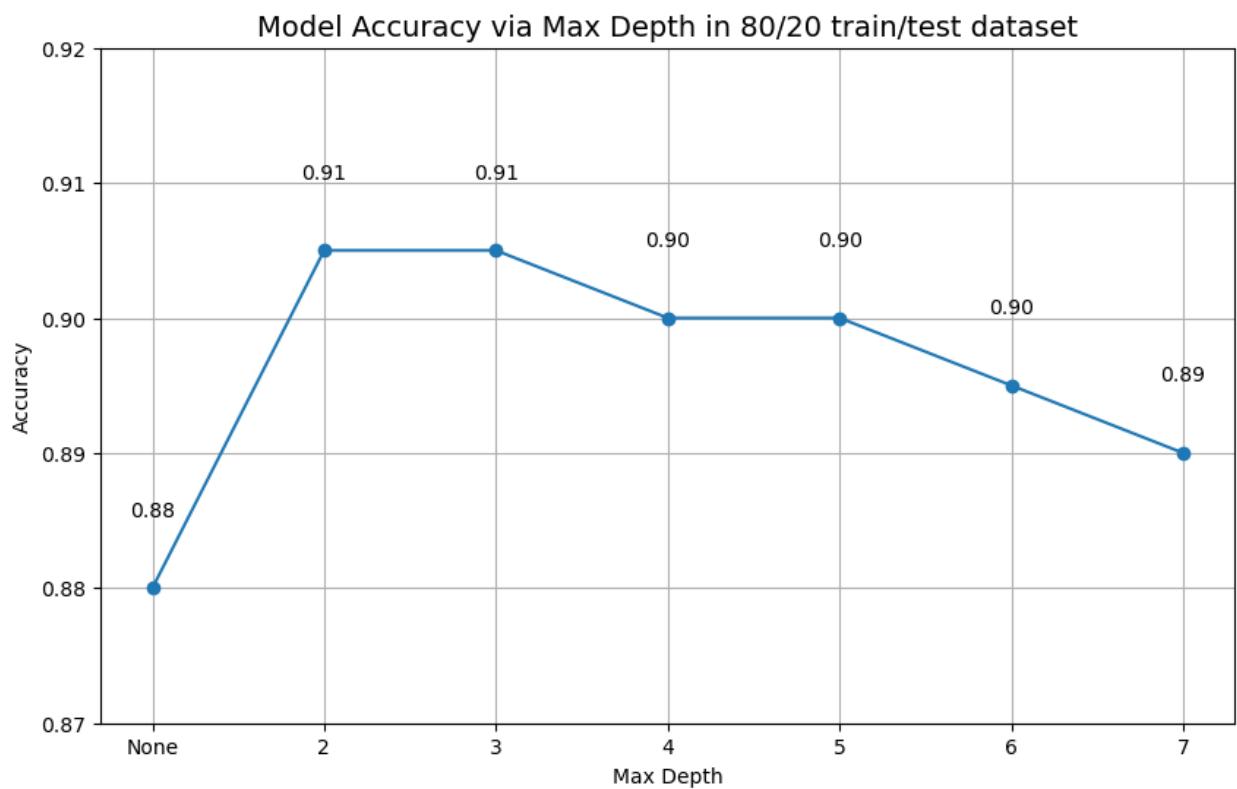
Max depth = 6

Đường dẫn: Source\Addition\Image7Depths\DT_6.png

Max depth = 7

Đường dẫn: Source\Addition\Image7Depths\DT_7.png

So sánh các độ sâu



Insight:

- Hiệu suất tổng thể: Độ chính xác (Accuracy) của mô hình dao động từ 0.88 đến 0.91, cho thấy hiệu suất phân loại khá ổn định với các giá trị khác nhau của Max Depth.
- Ảnh hưởng của Max Depth:
 - o Khi Max_depth = None (cây không giới hạn độ sâu), Accuracy đạt 0.88, thấp nhất trong các giá trị. Điều này cho thấy cây có thể bị overfitting, dẫn đến hiệu suất giảm trên tập kiểm tra.
 - o Khi Max Depth được giới hạn:
 - Max_depth = 2 và 3: Accuracy đạt giá trị cao nhất 0.91, cho thấy cây có độ sâu vừa phải giúp mô hình tổng quát tốt hơn.
 - Max_depth = 4, 5, 6, 7: Accuracy giảm nhẹ xuống còn 0.89-0.90, có thể do cây bắt đầu học quá chi tiết từ dữ liệu huấn luyện (overfitting nhẹ).
- Xu hướng:
 - o Max Depth thấp (2-3): Mô hình đơn giản hơn, nhưng đạt hiệu suất tối ưu (0.91). Điều này cho thấy các đặc trưng trong dữ liệu có thể được phân loại hiệu quả chỉ với một số mức phân chia giới hạn.
 - o Max Depth cao (4-7): Accuracy không tăng thêm mà giảm nhẹ, chứng tỏ việc tăng độ sâu không mang lại lợi ích, có thể dẫn đến học quá mức dữ liệu (overfitting).
- Kết luận:
 - o Max_depth = 2 hoặc 3 là lựa chọn tối ưu, với độ chính xác cao nhất (0.91) và mô hình đơn giản, dễ giải thích.

3. So sánh 3 dataset

- Đặc điểm bộ dữ liệu
 - o Breast Cancer dataset có 30 đặc trưng, kích thước mẫu là 569 và có 2 phân lớp. Số lượng lớp nhỏ giúp cây quyết định dễ dàng phân biệt giữa các lớp, dẫn đến độ chính xác, độ chính xác và độ nhớ cao.
 - o Wine Quality dataset có 11 đặc trưng, kích thước mẫu là 4898 và có 10 phân lớp được chia thành 3 nhóm. Số lượng lớp lớn hơn và đặc trưng ít hơn làm tăng độ phức tạp của nhiệm vụ phân loại, khiến mô hình khó đạt được hiệu suất cao trên tất cả các lớp, dẫn đến độ chính xác, độ chính xác và độ nhớ thấp hơn.
 - o Car Purchase dataset có 3 đặc trưng, kích thước mẫu là 1000 và có 2 phân lớp. Nhiệm vụ phân loại nhị phân với số đặc trưng nhỏ giúp cây quyết định dễ dàng phân biệt giữa các lớp, dẫn đến hiệu suất cao.
- Kết quả phân loại
 - o Breast Cancer: Mô hình cho thấy độ chính xác và độ nhớ cao cho cả hai lớp. Số lượng lớp nhỏ và tầm quan trọng của việc xác định chính xác cả các trường hợp ác tính và lành tính có thể đã góp phần vào các chỉ số cao này. Đặc điểm của bộ dữ liệu cho phép cây quyết định hoạt động tốt trong cả độ chính xác và độ nhớ.
 - o Wine Quality: Độ chính xác và độ nhớ thấp hơn cho một số lớp. Tính chất đa lớp của bộ dữ liệu và các mức chất lượng rượu khác nhau khiến mô hình khó đạt được độ chính xác và độ nhớ cao trên tất cả các lớp. Kích thước mẫu lớn hơn và số lượng đặc trưng ít hơn cũng gây khó khăn trong việc phân loại chính xác từng mức chất lượng.
 - o Car Purchase: Mô hình đạt được độ chính xác và độ nhớ cao cho cả hai lớp. Nhiệm vụ phân loại nhị phân và các đặc điểm của bộ dữ liệu (số lượng đặc trưng và mẫu vừa phải) cho phép cây quyết định phân biệt hiệu quả giữa mua và không mua, dẫn đến độ chính xác và độ nhớ cao.

4. Phân tích các đặc điểm ảnh hưởng đến cây quyết định của các dataset

4.1 Số lượng lớp

- Số lượng lớp nhỏ (như trong bộ dữ liệu Breast Cancer và Car Purchase) giúp cây quyết định dễ dàng phân biệt giữa các lớp, dẫn đến hiệu suất cao. Ngược lại, số lượng lớp lớn (như trong bộ dữ liệu Wine Quality) làm tăng độ phức tạp của nhiệm vụ phân loại, dẫn đến hiệu suất thấp hơn.
- Khi số lượng lớp tăng, cây quyết định cần phân chia phức tạp hơn để xử lý sự đa dạng giữa các lớp. Điều này dẫn đến:
 - o Tăng độ sâu của cây để bao phủ tất cả các lớp.
 - o Nguy cơ overfitting nếu số mẫu không đủ lớn để hỗ trợ việc phân chia.
 - o Nếu phân phối các lớp không cân bằng, các lớp nhỏ thường bị mô hình bỏ sót, làm giảm hiệu suất (Recall và Precision).

4.2 Số đặc trưng

- Số lượng đặc trưng vừa phải (như trong bộ dữ liệu Breast Cancer và Car Purchase) giúp cây quyết định học các mẫu mà không bị overfitting, góp phần vào hiệu suất cao. Số lượng đặc trưng ít hơn (như trong bộ dữ liệu Wine Quality) có thể không đủ để mô hình học được các mẫu phức tạp, dẫn đến hiệu suất thấp hơn.
- Số lượng đặc trưng cao có thể làm tăng độ phức tạp của cây:
 - o Cây cần tìm kiêm và so sánh trên nhiều thuộc tính, làm tăng thời gian huấn luyện.
 - o Có thể dẫn đến overfitting nếu cây chọn những đặc trưng không liên quan.
- Số đặc trưng thấp có thể dẫn đến underfitting, vì cây không có đủ thông tin để phân biệt giữa các lớp.

4.3 Kích thước mẫu

- Số lượng mẫu lớn:
 - o Cải thiện độ chính xác vì cây có thể học được nhiều mẫu hơn, dẫn đến khả năng tổng quát tốt hơn.
 - o Tuy nhiên, thời gian huấn luyện và kích thước cây sẽ tăng.
- Số lượng mẫu nhỏ:
 - o Dễ gây overfitting, vì cây có xu hướng học quá kỹ từ dữ liệu huấn luyện.
 - o Hiệu suất giảm khi cây không thể tổng quát tốt trên tập kiểm tra.
- Kích thước mẫu vừa phải (như trong bộ dữ liệu Breast Cancer và Car Purchase) giúp mô hình học được các mẫu mà không bị overfitting, dẫn đến hiệu suất cao. Kích thước mẫu lớn hơn (như trong bộ dữ liệu Wine Quality) làm tăng độ phức tạp của nhiệm vụ phân loại, dẫn đến hiệu suất thấp hơn.

4.4 Kết luận

- Số lượng lớp: Số lượng lớp nhỏ giúp cây quyết định dễ dàng phân biệt giữa các lớp, dẫn đến hiệu suất cao. Ngược lại, số lượng lớp lớn làm tăng độ phức tạp của nhiệm vụ phân loại, dẫn đến hiệu suất thấp hơn.
- Số lượng đặc trưng: Số lượng đặc trưng vừa phải giúp cây quyết định học các mẫu mà không bị overfitting, góp phần vào hiệu suất cao. Số lượng đặc trưng ít hơn có thể không đủ để mô hình học được các mẫu phức tạp, dẫn đến hiệu suất thấp hơn.
- Kích thước mẫu: Kích thước mẫu vừa phải giúp mô hình học được các mẫu mà không bị quá khớp, dẫn đến hiệu suất cao. Kích thước mẫu lớn hơn giúp mô hình học được nhiều mẫu hơn, nhưng cũng làm tăng độ phức tạp của nhiệm vụ phân loại, dẫn đến hiệu suất thấp hơn nếu mô hình quá phức tạp so với lượng dữ liệu.

TÀI LIỆU THAM KHẢO

- [1] scikit-learn, "scikit-learn," [Online]. Available: <https://scikit-learn.org/1.5/modules/tree.html>.