

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN



---

## DECISION TREE

Course: CS14003 – Cơ Sở Trí Tuệ Nhân Tạo

---

Sinh viên:

Mạch Quốc Tân  
Nguyễn Thành Dâng  
Nguyễn Thị Khánh Linh  
Trương Thành Đạt

Giáo viên hướng dẫn:

Bùi Duy Đăng  
Lê Nhựt Nam  
Nguyễn Thanh Tình  
Huỳnh Lâm Hải Đăng

# Mục lục

<b>1 Thông tin chung</b>	<b>3</b>
1.1 Thông tin đồ án . . . . .	3
1.2 Mô tả . . . . .	3
1.3 Thông tin thành viên . . . . .	4
1.4 Bảng phân chia công việc . . . . .	4
<b>2 Đánh giá đồ án</b>	<b>5</b>
<b>3 Một số khái niệm cần biết</b>	<b>6</b>
<b>4 Phân tích dữ liệu</b>	<b>8</b>
4.1 Tập dữ liệu Breast Cancer . . . . .	8
4.1.1 Chuẩn bị dữ liệu . . . . .	8
4.1.2 Xây dựng mô hình cây quyết định (Decision Tree) . . . . .	10
4.1.3 Đánh giá mô hình Decision Tree . . . . .	15
4.1.4 Đánh giá độ chính xác theo độ sâu . . . . .	19
4.2 Tập dữ liệu Wine Quality . . . . .	22
4.2.1 Chuẩn bị dữ liệu . . . . .	22
4.2.2 Xây dựng mô hình cây quyết định (Decision Tree) . . . . .	24
4.2.3 Đánh giá mô hình Decision Tree . . . . .	28
4.2.4 Đánh giá độ chính xác theo độ sâu . . . . .	33
4.3 Tập dữ liệu bổ sung (Age Prediction) . . . . .	36
4.3.1 Chuẩn bị dữ liệu . . . . .	36
4.3.2 Xây dựng cây quyết định (Decision tree) . . . . .	37
4.3.3 Đánh giá mô hình Decision Tree . . . . .	42
4.3.4 Đánh giá độ chính xác theo độ sâu . . . . .	46
4.4 So sánh giữa các tập dữ liệu . . . . .	49
4.4.1 Tổng quan . . . . .	49

4.4.2 Độ chính xác tổng thể (Accuracy) . . . . .	49
4.4.3 Hiệu suất theo từng lớp (F1-score) . . . . .	50
4.4.4 Ảnh hưởng của độ sâu cây . . . . .	50
4.5 Kết luận và đề xuất . . . . .	51
<b>5 Tham khảo</b>	<b>51</b>

# 1 Thông tin chung

## 1.1 Thông tin đồ án

- Tên học phần: Cơ sở trí tuệ nhân tạo
- Đồ án thực hiện: Decision Tree
- Source code: [Github](#)

## 1.2 Mô tả

Trong đồ án này, ta thực hiện tìm hiểu và áp dụng mô hình cây quyết định (Decision Tree) để giải quyết các bài toán phân loại trên dữ liệu thực tế. Đây là một trong những phương pháp đơn giản nhưng hiệu quả trong học có giám sát.

Quy trình thực hiện:

- Chuẩn bị dữ liệu từ ba tập: Breast Cancer Wisconsin, Wine Quality, và một tập dữ liệu bổ sung do nhóm tự chọn.
- Tiền xử lý dữ liệu, chia thành tập huấn luyện và kiểm thử với nhiều tỉ lệ khác nhau (40/60, 60/40, 80/20, 90/10).
- Huấn luyện mô hình cây quyết định sử dụng thư viện scikit-learn, trực quan hóa cây và đánh giá hiệu quả mô hình qua các chỉ số: độ chính xác, báo cáo phân loại, và ma trận nhầm lẫn.
- Thủ nghiệm thay đổi độ sâu của cây để phân tích ảnh hưởng đến độ chính xác của mô hình.
- Sau khi hoàn tất các phần thực nghiệm, nhóm tiến hành so sánh hiệu suất mô hình trên ba bộ dữ liệu và đưa ra nhận xét về ảnh hưởng của số lượng lớp, số lượng thuộc tính và kích thước tập dữ liệu đến kết quả phân loại.

### 1.3 Thông tin thành viên

MSSV	Họ và tên	Email
23127115	Mạch Quốc Tân	mqtan23@clc.fitus.edu.vn
23127334	Nguyễn Thành Dâng	ntdang23@clc.fitus.edu.vn
23127344	Trương Thành Đạt	ttdat23@clc.fitus.edu.vn
23127082	Nguyễn Thị Khánh Linh	ntklinh23@clc.fitus.edu.vn

Bảng 1: Bảng thông tin thành viên

### 1.4 Bảng phân chia công việc

Công việc	Phụ trách	Trạng thái	%
Thực thi mã nguồn xây dựng Desision Tree trên ba bộ dữ liệu	Quốc Tân	Hoàn Thành	100%
Phân tích và viết báo cáo về bộ dữ liệu Breast Cancer và so sánh các bộ dữ liệu	Thành Dâng	Hoàn Thành	100%
Phân tích và viết báo cáo về bộ dữ liệu Wine Quality	Khánh Linh	Hoàn Thành	100%
Phân tích và viết báo cáo về bộ dữ liệu bổ sung	Thành Đạt	Hoàn Thành	100%

Bảng 2: Bảng phân chia công việc

## 2 Đánh giá đồ án

Trong đồ án này, nhóm chúng em đã học hỏi được rất nhiều thông qua việc tìm hiểu, cài đặt mô hình Decision Tree và phân tích mô hình ấy trên các bộ dữ liệu khác nhau. Qua đó, nhóm chúng em thấy rằng nhóm đã hoàn thành tốt các yêu cầu đề ra, tự đánh giá với số điểm là **10/10**, cụ thể ở các tiêu chí sau:

**Chuẩn bị dữ liệu:** Việc chuẩn bị dữ liệu dựa trên 2 bộ dữ liệu được cung cấp, chúng em đã tiến hành định dạng các labels một cách trực quan hơn thông qua việc ghi rõ hơn labels của tập dữ liệu Breast Cancer, gom nhóm các labels thành 3 labels chính. Ngoài ra chúng em đã tìm thêm và thực thi trên bộ dữ liệu dự đoán độ tuổi thông qua các chỉ số cơ thể. Dữ liệu được tiến hành phân chia theo tỉ lệ yêu cầu, đảm bảo được tính ngẫu nhiên và tính đúng đắn trong tỉ lệ loại, được thể hiện thông qua việc trực quan hóa các tập dữ liệu được chia. Thực thi hoàn chỉnh ở ba bộ dữ liệu.

**Thực thi Decision Tree để phân loại:** Tìm hiểu và sử dụng thư viện sklearn, chúng em đã thực thi thành công việc xây dựng Decision Tree với tiêu chí Information Gain và các thông số mặc định. Xây dựng đầy đủ trên các bộ dữ liệu với các tỉ lệ khác nhau. Và hiển thị đồ họa Decision Tree bởi sự hỗ trợ của thư viện Graphviz. Thực thi đầy đủ ở 3 bộ dữ liệu.

**Đánh giá hiệu suất của Decision Tree:** Với việc xây dựng thành công Decision Tree, áp dụng mô hình vào các dữ liệu với các tỉ lệ khác nhau, cho ra các số liệu cụ thể tương ứng, chúng em đã dựa vào đó để đánh giá mô hình với các tỉ lệ khác nhau, tìm hiểu các ý nghĩa của các thông số, qua đó đưa ra góc nhìn về độ hiệu quả của mô hình. Ngoài ra, chúng em đã thực hiện tạo ra các báo cáo số liệu bằng **classification report** và **confusion matrix** với sự hỗ trợ của các thư viện. Thông qua đó, chúng em đã phân tích chi tiết và đầy đủ các thông số để có thể đánh giá mô hình một cách tốt nhất. Thực hiện đầy đủ ở 3 bộ dữ liệu.

**Thực nghiệm về độ ảnh hưởng của chiều sâu của Decision Tree với độ chính xác:** Chúng em đã thực thi việc thử nghiệm xây dựng mô hình với giới hạn về độ sâu, qua đó thu nhận được các độ chính xác tương ứng sử dụng cho việc phân tích, ngoài ra hiển thị các Decision Tree bởi Graphviz và đưa ra bảng thông kê về độ sâu và độ chính xác tương ứng, điều này còn được thể hiện thông qua biểu đồ đường, mang đến góc nhìn trực quan hóa về sự thay đổi của độ sâu đến với độ chính xác. Chúng em đã dựa trên các thông số để phân tích rõ ràng và chi tiết về mối liên hệ giữa 2 đại lượng nêu trên. Thực hiện đầy đủ ở 3 bộ dữ liệu.

**So sánh các bộ dữ liệu:** Với các thông số đã phân tích phía trên, chúng em kết hợp và so sánh các tập dữ liệu với nhau để xem xét về chất lượng, tính hiệu quả của mô hình trên các bộ dữ liệu, điều này được chúng em phân tích rõ ràng và chi tiết ở phần báo cáo.

**Các notebook được cấu trúc và định dạng tốt:** Các notebook hỗ trợ việc thực thi được chúng em tổ chức gọn gàng, ngoài ra thêm thắt các icon tạo cho notebook thêm phần sống động hơn, giải thích rõ ràng các bước thực thi xây dựng mô hình.

### 3 Một số khái niệm cần biết

Báo cáo phân loại (Classification report) cung cấp một số chỉ số quan trọng để đánh giá hiệu suất của một mô hình phân loại. Các chỉ số đó bao gồm:

- **Precision:** mức độ chính xác của mô hình hay chính là tỷ số giữa số lần gán nhãn đúng cho mô hình với tổng số nhãn đã được gán.

$$\text{Precision} = \frac{\sum \text{True Positive (TP)}}{\sum \text{True Positive (TP)} + \text{False Positive (FP)}} \quad (1)$$

- **True Positive (TP):** Số mẫu được dự đoán đúng là A (thực tế là A và mô hình dự đoán A)
- **False Positive (FP):** Số mẫu được dự đoán sai là A (thực tế là B nhưng mô hình dự đoán A)

- **Recall:** là độ bao phủ, hay chính là tỷ lệ mẫu thuộc một lớp nào đó được mô hình dự đoán đúng, so với tổng số mẫu thực sự thuộc lớp đó.

$$\text{Recall} = \frac{\sum \text{True Positive (TP)}}{\sum \text{True Positive (TP)} + \text{False Negative (FN)}} \quad (2)$$

- **True Positive (TP):** Số mẫu được dự đoán đúng là A (thực tế là A và mô hình dự đoán A)
- **False Negative (FN):** Số mẫu được dự đoán sai là B (thực tế là A nhưng mô hình dự đoán B)

- **F1-Score:** chỉ số giúp bạn đánh giá hiệu quả mô hình phân loại, bằng cách kết hợp Precision và Recall. Ta biết rằng khi precision cao nghĩa là mô hình có ít nhầm lẫn, Khi Recall cao nghĩa là mô hình ít bỏ sót, nếu một trong 2 thấp thì mô hình đó vẫn chưa ổn. Vì vậy F1-Score sẽ là giá trị trung bình điều hòa của Percision và Recall, nếu một trong thấp thì F1-Score cũng thấp theo. Nhờ đó, F1-Score là một thước đo đặc biệt hữu ích trong các bài toán phân loại có dữ liệu mất cân đối, hoặc khi ta cần đánh giá mô hình không chỉ chính xác mà còn không bỏ sót các trường hợp quan trọng.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

- **Support:** Số lượng thực sự của mỗi label.

Ma trận nhầm lẩn (confusion matrix) là một bảng thể hiện cách mô hình phân loại hoạt động trên tập dữ liệu kiểm tra, bằng cách hiển thị số lượng dự đoán đúng và sai dựa trên các thành phần:

		Positive	TP	FN
	<b>Actual value</b>			
Negative		Positive	FP	TN
				<b>Negative</b>
		<b>Predicted value</b>		

Hình 1: Ma trận nhầm lẩn (Confusion matrix)

- True Positives (TP): Dương tính đúng, những mẫu thực sự thuộc lớp dương tính và mô hình cũng dự đoán đúng dương tính.
- True Negatives (TN): Âm tính đúng, những mẫu thực sự thuộc lớp âm tính và mô hình cũng dự đoán đúng âm tính.
- False Positives (FP): Dương tính sai, những mẫu thực sự thuộc lớp âm tính nhưng mô hình lại dự đoán là dương tính.
- False Negatives (FN): Âm tính sai, những mẫu thực sự thuộc lớp dương tính nhưng mô hình dự đoán âm tính.

## 4 Phân tích dữ liệu

### 4.1 Tập dữ liệu Breast Cancer

#### 4.1.1 Chuẩn bị dữ liệu

Bộ dữ liệu **Chẩn đoán Ung thư Vú Wisconsin** nhằm phục vụ cho mục đích phân loại khối u là lành tính (**B**) hoặc ác tính (**M**) dựa trên 30 Features được trích từ hình ảnh. Tổng cộng có 569 mẫu dữ liệu, mỗi mẫu đều được gán nhãn là **M** (Malignant – Ác tính) hoặc **B** (Benign – Lành tính) đồng thời mỗi mẫu được mô tả bởi 30 đặc trưng bao gồm các giá trị bán kính, chu vi, diện tích, độ mịn, độ đặc, v.v.

Hàm `load_data()` sẽ được sử dụng để tải dữ liệu và chuẩn bị cho quá trình phân chia thành tập train và tập test. Với cả hai tập đều là với bộ dữ liệu ban đầu, các dữ liệu features và label sẽ được xáo trộn theo cùng một thứ tự bằng hàm `shuffle()` để giữ nguyên mối liên hệ giữa Features và Label.

Sau khi xáo trộn Features và Labels, bộ dữ liệu sẽ được chia thành tập huấn luyện (train) và tập kiểm tra (test) theo các tỷ lệ cho trước. Các tỷ lệ huấn luyện/kiểm tra khác nhau được sử dụng gồm: 40/60, 60/40, 80/20 và 90/10.

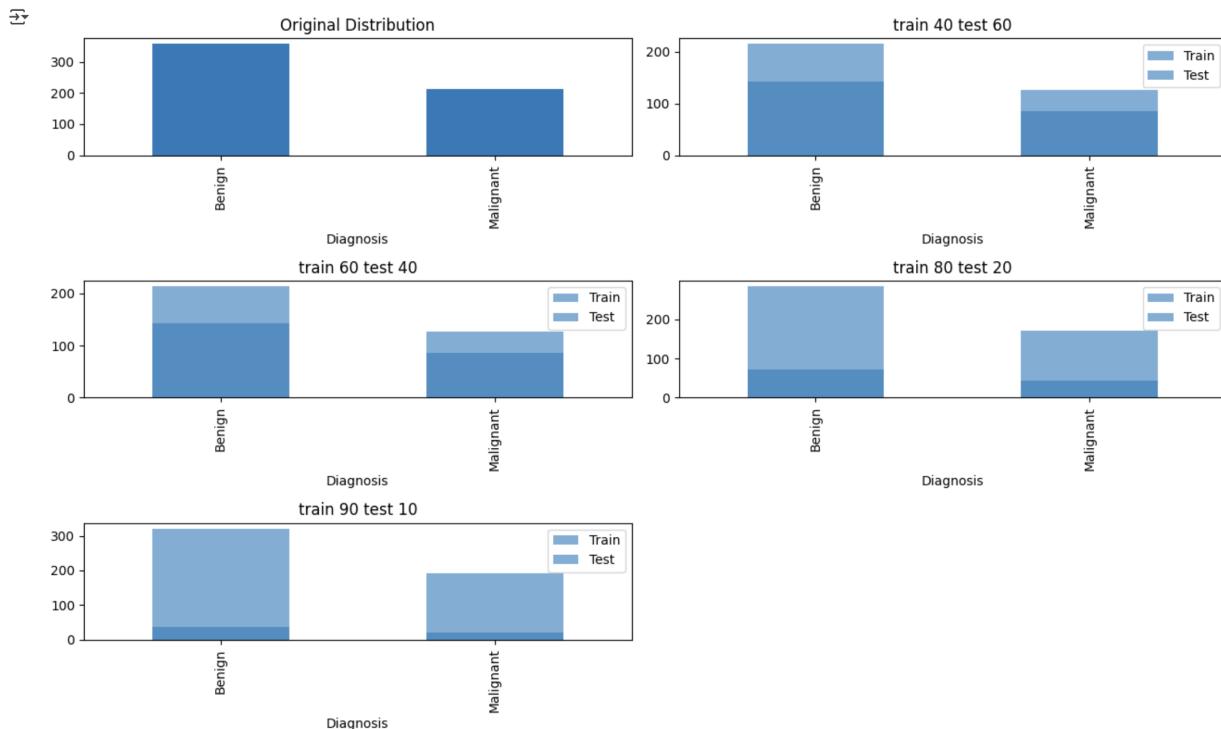
Sau khi chia dữ liệu bằng hàm `create_stratified_splits()` với các tham số:

- **stratify**: để giữ nguyên phân bố nhãn, tham số đảm bảo rằng tỷ lệ phân bố Labels trên tập train và tập test là không đổi.
- **random\_state=42**: để đảm bảo kết quả nhất quán giữa các lần chạy

các bộ dữ liệu này vào một danh sách gọi là `splits`. Mỗi phần tử trong danh sách này là một bộ 4 phần tử bao gồm:

- **feature\_train**: Tập mẫu huấn luyện (không bao gồm nhãn)
- **feature\_test**: Tập mẫu kiểm tra (cùng cấu trúc với `feature_train`)
- **label\_train**: Nhãn tương ứng với các mẫu trong `feature_train`
- **label\_test**: Nhãn tương ứng với các mẫu trong `feature_test`

Như vậy, sau khi hoàn tất quá trình chia dữ liệu, với mỗi tỷ lệ chúng ta sẽ có 4 tập con, tổng cộng là 16 bộ dữ liệu.



Hình 2: Tỷ lệ giữa tập train và tập test

Sau khi hoàn thành việc chia và chuẩn bị bộ dữ liệu ở các tỷ lệ khác nhau, chúng ta sẽ tiến hành thể hiện sự phân bố dữ liệu giữa các lớp (**class**), tập trung vào:

- Tập huấn luyện (training set)
- Tập kiểm tra (testing set)

Quá trình trực quan hóa (**Visualization**) sẽ sử dụng biểu đồ cột chồng để so sánh số lượng mẫu của Label "Benign" và "Malinant" trong tập dữ liệu train và test với các tỷ lệ chia dữ liệu khác nhau.

Việc trực quan hóa này mang lại lợi ích chính:

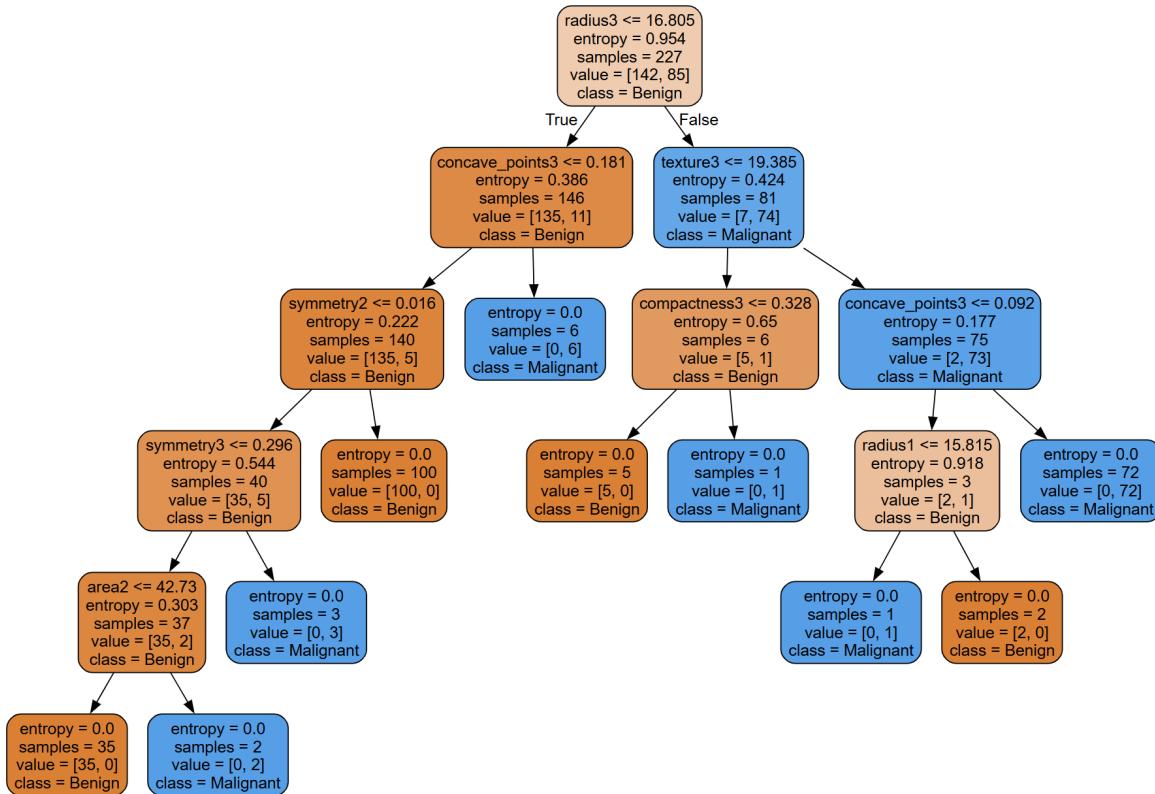
- Dễ dàng nhìn thấy được sự phân bố của các label trong mỗi tập dữ liệu.
- Cung cấp cái nhìn về ảnh hưởng của quá trình chia dữ liệu lên tập dữ liệu ban đầu.

#### 4.1.2 Xây dựng mô hình cây quyết định (Decision Tree)

Hàm `evaluate_tree()` được sử dụng để tạo và huấn luyện mô hình cây quyết định (**Decision Tree Classifier**) trên một tập dữ liệu đã được chia sẵn thành tập train và test. Mô hình được xây dựng sử dụng tiêu chí **entropy** nhằm lựa chọn thuộc tính phân nhánh dựa trên chỉ số **Information Gain**. Ngoài ra, tham số `max_depth` có thể được truyền vào để giới hạn độ sâu của cây, giúp kiểm soát mức độ phức tạp của mô hình.

Sau khi khởi tạo, mô hình sẽ được huấn luyện trên tập dữ liệu huấn luyện bằng phương thức `fit()`, giúp cây quyết định học được mối quan hệ giữa đặc trưng và nhãn trong dữ liệu. Hàm trả về mô hình cây quyết định đã được huấn luyện, sẵn sàng để sử dụng trong các bước đánh giá hoặc dự đoán tiếp theo.

Để thể hiện rõ cấu trúc và cách thức hoạt động của mô hình, mỗi cây quyết định sau khi được huấn luyện sẽ được trực quan hóa (**Visualization**) bằng thư viện **graphviz**. Cây quyết định được hiển thị dưới dạng sơ đồ phân nhánh, trong đó mỗi Node biểu diễn một điều kiện phân tách dựa trên một features cụ thể, các nhánh thể hiện hướng đi tùy thuộc vào việc điều kiện đó đúng hay sai, và các nút lá thể hiện lớp Label mà mô hình dự đoán. Thông tin hiển thị tại mỗi nút bao gồm tên feature, giá trị threshold, chỉ số entropy, số lượng mẫu và lớp label.

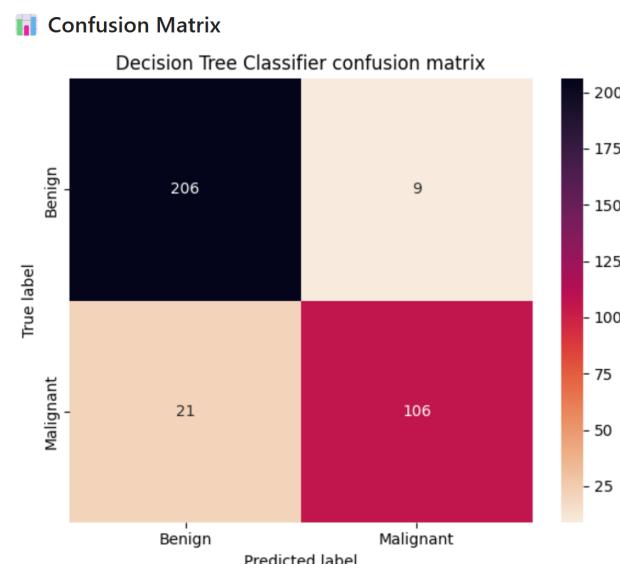


Hình 3: Decision Tree với tỷ lệ 40/60

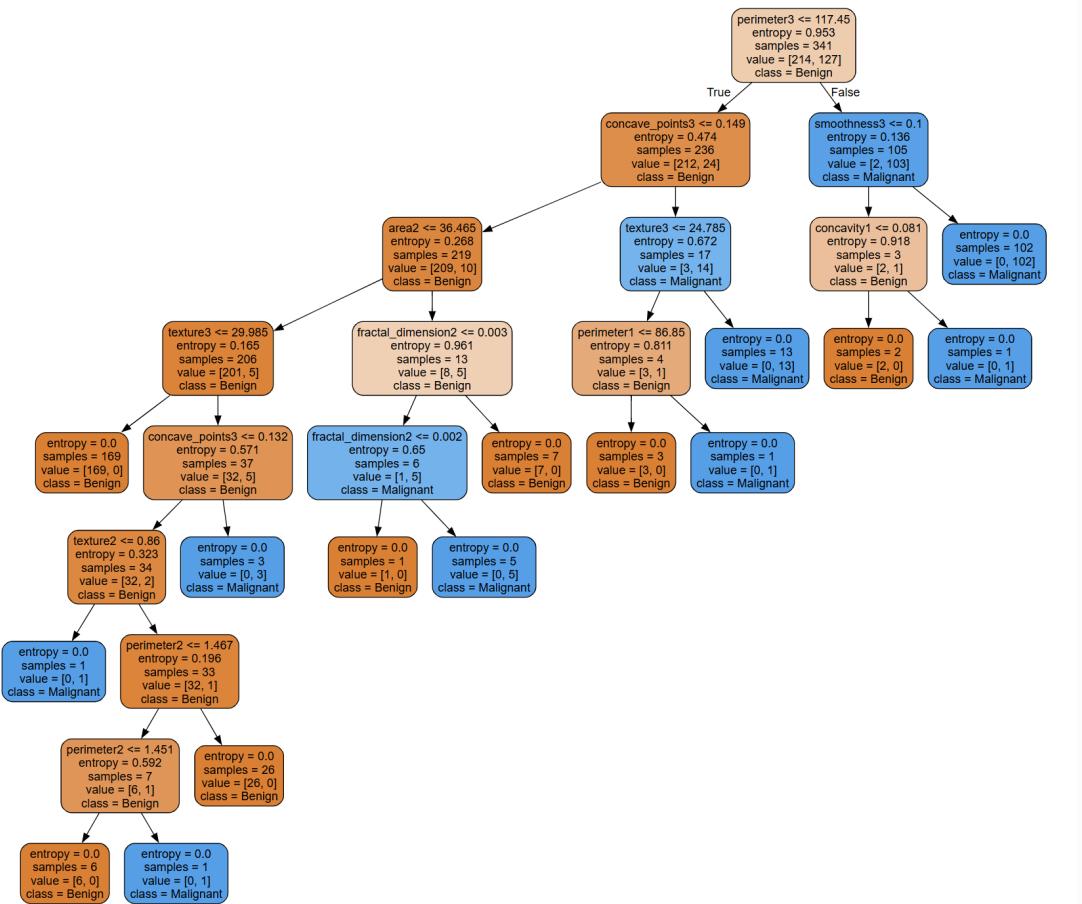
#### Classification Report

	precision	recall	f1-score	support
Benign	0.91	0.96	0.93	215
Malignant	0.92	0.83	0.88	127
accuracy			0.91	342
macro avg	0.91	0.90	0.90	342
weighted avg	0.91	0.91	0.91	342
Accuracy:	0.91			

Hình 4: Phân loại với tỷ lệ 40/60



Hình 5: Ma trận nhầm lẫn với tỷ lệ 40/60



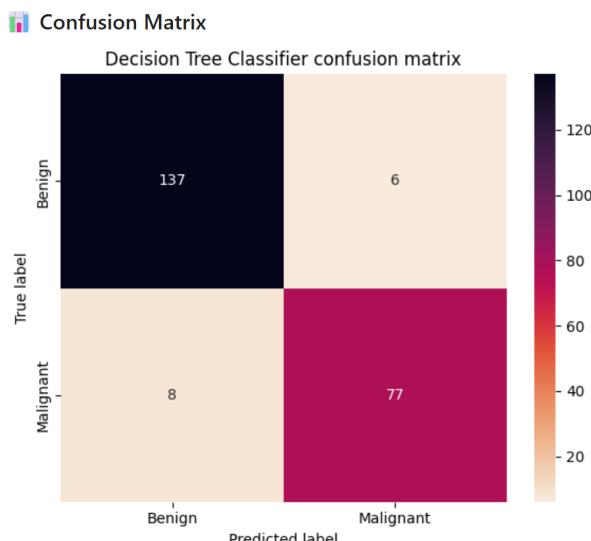
Hình 6: Decision Tree với tỷ lệ 60/40

## Classification Report

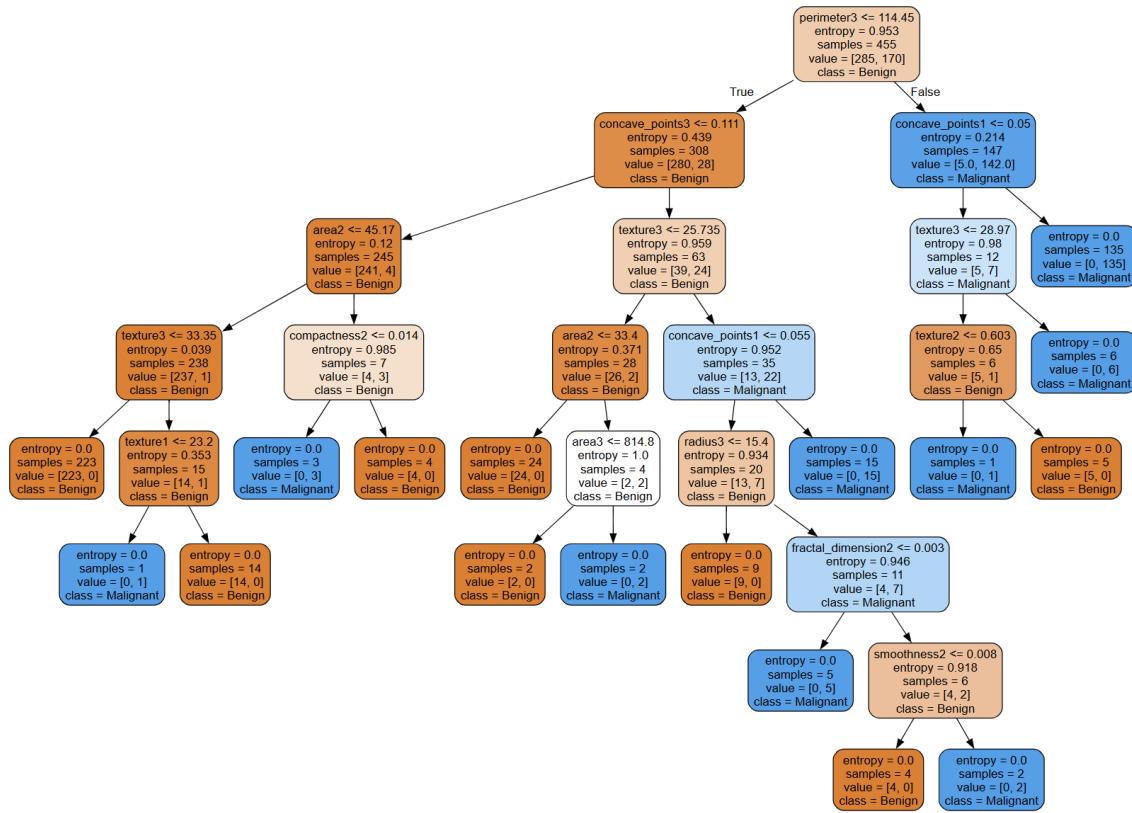
	precision	recall	f1-score	support
Benign	0.94	0.96	0.95	143
Malignant	0.93	0.91	0.92	85
accuracy			0.94	228
macro avg	0.94	0.93	0.93	228
weighted avg	0.94	0.94	0.94	228

Accuracy: 0.94

Hình 7: Phân loại với tỷ lệ 60/40



Hình 8: Ma trận nhầm lẩn với tỷ lệ 60/40



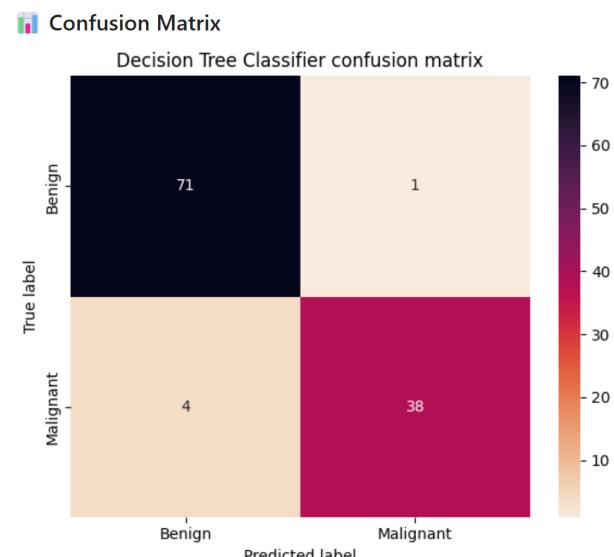
Hình 9: Decision Tree với tỷ lệ 80/20

### Classification Report

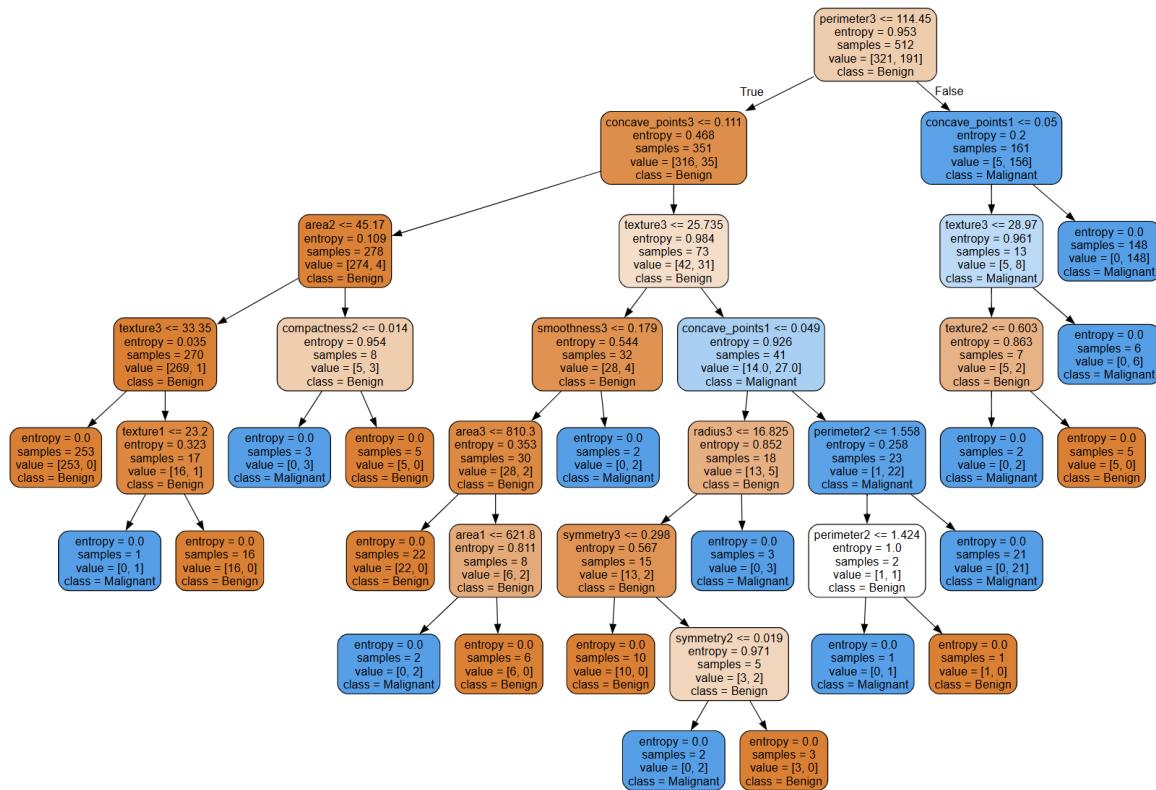
	precision	recall	f1-score	support
Benign	0.95	0.99	0.97	72
Malignant	0.97	0.90	0.94	42
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

Accuracy: 0.96

Hình 10: Phân loại với tỷ lệ 80/20



Hình 11: Ma trận nhầm lẫn với tỷ lệ 80/20



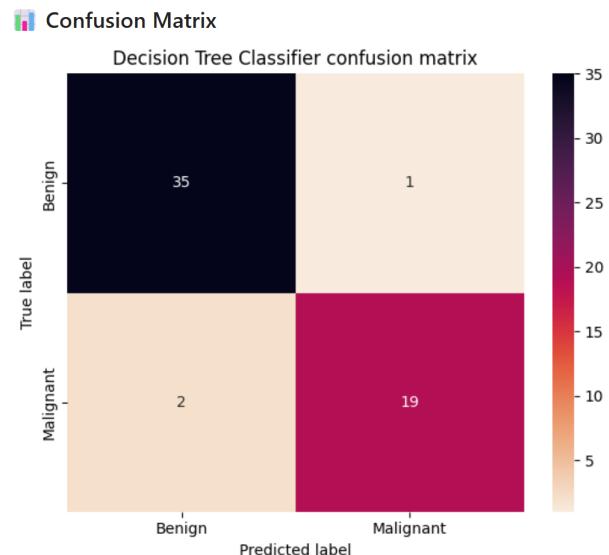
Hình 12: Decision Tree với tỷ lệ 90/10

## Classification Report

	precision	recall	f1-score	support
Benign	0.95	0.97	0.96	36
Malignant	0.95	0.90	0.93	21
accuracy			0.95	57
macro avg	0.95	0.94	0.94	57
weighted avg	0.95	0.95	0.95	57

Accuracy: 0.95

Hình 13: Phân loại với tỷ lệ 90/10



Hình 14: Ma trận nhầm lẫn với tỷ lệ 90/10

#### 4.1.3 Đánh giá mô hình Decision Tree

Kết luận mô hình theo tỷ lệ

Mô hình tỷ lệ 40/60

Classification Report				
	precision	recall	f1-score	support
Benign	0.91	0.96	0.93	215
Malignant	0.92	0.83	0.88	127
accuracy			0.91	342
macro avg	0.91	0.90	0.90	342
weighted avg	0.91	0.91	0.91	342
Accuracy: 0.91				

Hình 15: Classification report tỷ lệ 40/60

Độ chính xác của mô hình (Accuracy) là 91% cho thấy khả năng phân loại tốt trên hầu hết các lớp dữ liệu, tức là mô hình gần đúng nhãn cho phần lớn mẫu thử. Đây là một kết quả tốt, tuy nhiên mô hình vẫn cần được xem xét kỹ để cải thiện thêm – đặc biệt trong những trường hợp mà precision hoặc recall có thể chưa cao.

Trong lớp "Benign"(B):

- Mô hình đạt precision cao 91% nghĩa là số lượng mẫu bị dự đoán sai thành “Benign” là rất ít (ít false positive).
- Recall là 96%, nghĩa là mô hình đã dự đoán đúng 93% số lượng mẫu thật sự thuộc lớp “Benign”, chỉ bỏ sót khoảng 4%.
- F1-Score là 93%, cho thấy mô hình có sự cân bằng tốt giữa precision và recall, từ đó khẳng định hiệu suất dự đoán ổn định và chính xác cho lớp “Benign”.

Trong lớp "Malignant"(M):

- Precision là 92% cao hơn so với percision của (Benign) có nghĩa là tỷ lệ dự đoán chính xác của (Malignant) cao hơn (Benign) trong một vài trường hợp.
- Recall của lớp này thấp hơn một chút so với lớp (Benign), tuy nhiên vẫn ở mức cao, chỉ chênh lệch nhẹ – điều này cho thấy vẫn còn một số trường hợp bị dự đoán bỏ sót.
- F1-Score đạt 90.77%, cho thấy mô hình vẫn giữ được sự cân bằng tốt giữa precision và recall, và thể hiện hiệu suất ổn định trong việc dự đoán các ca ác tính (malignant).

## Mô hình tỷ lệ 60/40

Classification Report				
	precision	recall	f1-score	support
Benign	0.94	0.96	0.95	143
Malignant	0.93	0.91	0.92	85
accuracy			0.94	228
macro avg	0.94	0.93	0.93	228
weighted avg	0.94	0.94	0.94	228
Accuracy: 0.94				

Hình 16: Classification report tỷ lệ 60/40

Độ chính xác của mô hình (Accuracy) là 94% cho thấy khả năng phân loại tốt trên hầu hết các lớp dữ liệu, tức là mô hình gán đúng nhãn cho phần lớn mẫu thử.

Trong lớp "Benign"(B):

- Mô hình đạt precision cao 94% nghĩa là số lượng mẫu bị dự đoán sai thành “Benign” là rất ít (ít false positive).
- Recall là 96%, nghĩa là mô hình đã dự đoán đúng 93% số lượng mẫu thật sự thuộc lớp “Benign”, chỉ bỏ sót khoảng 4%.
- F1-Score là 95%, cho thấy mô hình có sự cân bằng tốt giữa precision và recall, từ đó khẳng định hiệu suất dự đoán ổn định và chính xác cho lớp “Benign”.

Trong lớp "Malignant"(M):

- Precision = 93% và Recall = 91 % của lớp này thấp hơn một chút so với lớp (Benign), tuy nhiên vẫn ở mức cao, chỉ chênh lệch nhẹ, điều này cho thấy vẫn còn một số trường hợp bị dự đoán sai.
- Mặc dù vậy, F1-Score vẫn đạt 92%, cho thấy mô hình vẫn giữ được sự cân bằng tốt giữa precision và recall, và thể hiện hiệu suất ổn định trong việc dự đoán các ca ác tính (malignant).

## Mô hình tỷ lệ 80/20

Classification Report				
	precision	recall	f1-score	support
Benign	0.95	0.99	0.97	72
Malignant	0.97	0.90	0.94	42
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114
Accuracy: 0.96				

Hình 17: Classification report tỷ lệ 80/20

Độ chính xác của mô hình (Accuracy) là 96% cho thấy khả năng phân loại tốt trên hầu hết các lớp dữ liệu, tức là mô hình gán đúng nhãn cho phần lớn mẫu thử.

Trong lớp "Benign"(B):

- Mô hình đạt precision cao 95% nghĩa là số lượng mẫu được đúng được dự đoán thành “Benign” rất nhiều và gần như là chính xác.
- Recall là 99%, nghĩa là mô hình đã dự đoán đúng 99% số lượng mẫu thật sự thuộc lớp “Benign”, chỉ bỏ sót khoảng 1%.
- F1-Score là 97%, cho thấy mô hình có sự cân bằng tốt giữa precision và recall, từ đó khẳng định hiệu suất dự đoán ổn định và chính xác cho lớp “Benign”.

Trong lớp "Malignant"(M):

- Precision là 97% cao hơn so với percision của (Benign) có nghĩa là tỷ lệ dự đoán chính xác của (Malignant) cao hơn (Benign) trong một vài trường hợp.
- Recall của lớp này thấp hơn so với lớp (Benign) gần 10% , tuy nhiên vẫn ở mức cao, chỉ chênh lệch nhẹ – điều này cho thấy vẫn còn một số trường hợp bị dự đoán sai.
- F1-Score đạt 94%, cho thấy mô hình vẫn giữ được sự cân bằng tốt giữa precision và recall, và thể hiện hiệu suất ổn định trong việc dự đoán các ca ác tính (malignant).

Mô hình tỷ lệ 90/10

Classification Report				
	precision	recall	f1-score	support
Benign	0.95	0.97	0.96	36
Malignant	0.95	0.90	0.93	21
accuracy			0.95	57
macro avg	0.95	0.94	0.94	57
weighted avg	0.95	0.95	0.95	57
Accuracy: 0.95				

Hình 18: Classification report tỷ lệ 90/10

Độ chính xác của mô hình (Accuracy) là 95% cho thấy khả năng phân loại tốt trên hầu hết các lớp dữ liệu, tức là mô hình gán đúng nhãn cho phần lớn mẫu thử.

Trong lớp "Benign"(B):

- Mô hình đạt precision cao 95% nghĩa là số lượng mẫu được đúng dự đoán thành “Benign” rất nhiều và gần như là chính xác.
- Recall là 97%, nghĩa là mô hình đã dự đoán đúng 97% số lượng mẫu thật sự thuộc lớp “Benign”, chỉ bỏ sót khoảng 3%.
- F1-Score là 96%, cho thấy mô hình có sự cân bằng tốt giữa precision và recall, từ đó khẳng định hiệu suất dự đoán ổn định và chính xác cho lớp “Benign”.

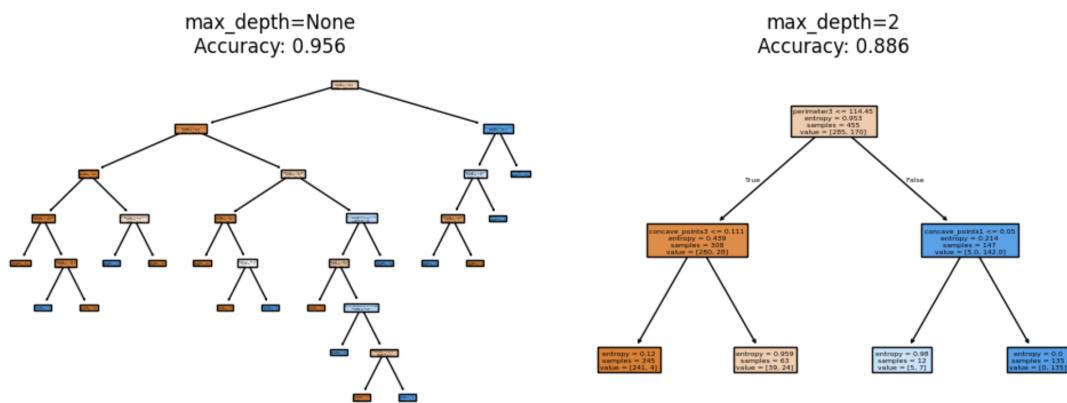
Trong lớp "Malignant"(M):

- Precision là 95% có tỷ lệ bằng so với percision của (Benign) có nghĩa là tỷ lệ dự đoán chính xác của (Malignant) tương đồng với tỷ lệ của (Benign).
- Recall của lớp này thấp hơn so với lớp (Benign) 7% , tuy nhiên vẫn ở mức cao, chỉ chênh lệch nhẹ – điều này cho thấy vẫn còn một số trường hợp bị dự đoán sai.
- F1-Score đạt 93%, cho thấy mô hình vẫn giữ được sự cân bằng tốt giữa precision và recall, và thể hiện hiệu suất ổn định trong việc dự đoán các ca ác tính (malignant).

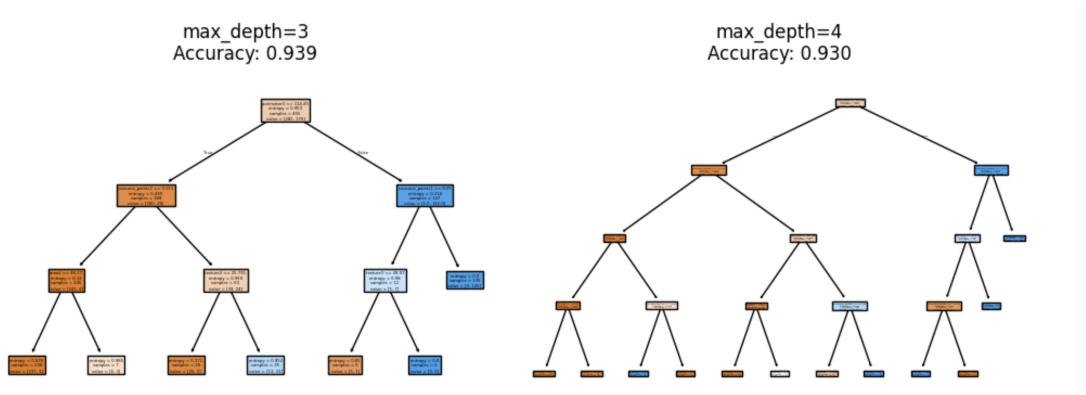
**Kết luận:** Trong bốn tỉ lệ chia dữ liệu huấn luyện và kiểm tra tương ứng với bốn mô hình cây quyết định khác nhau, mô hình sử dụng tỉ lệ 80/20 được đánh giá là phù hợp nhất. Lý do là vì mô hình này đạt độ chính xác khá cao (96%) cao nhất trong 4 trường hợp. Bên cạnh đó, việc sử dụng một tập train lớn giúp mô hình học được nhiều đặc điểm hơn từ tập dữ liệu, từ đó nâng cao giá trị và hiệu quả dự đoán của mô hình.

#### 4.1.4 Đánh giá độ chính xác theo độ sâu

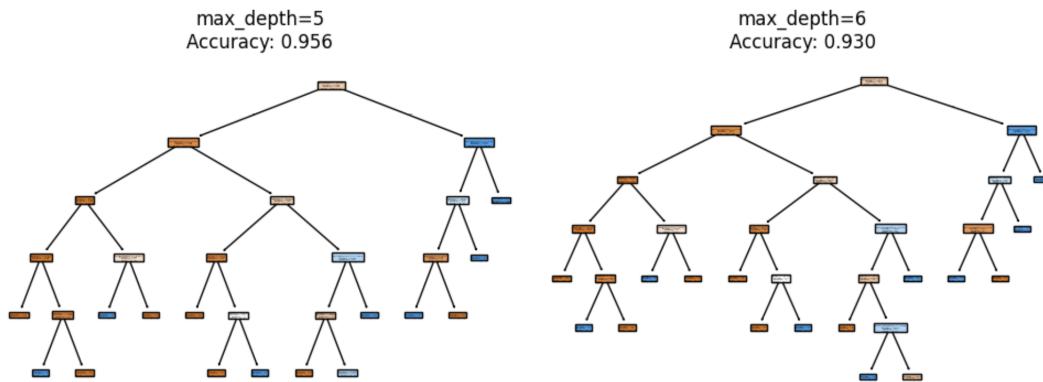
Sử dụng mô hình dự đoán Decision tree với tỷ lệ 80/20 ta sẽ thử nghiệm đối với mỗi độ sâu (Depth) khác nhau thì nó sẽ ảnh hưởng đến độ chính xác (Accuracy) của mô hình như thế nào. Ta sẽ lần lượt thử nghiệm với các độ sâu: 2, 3, 4, 5, 6, 7 và không giới hạn độ sâu.



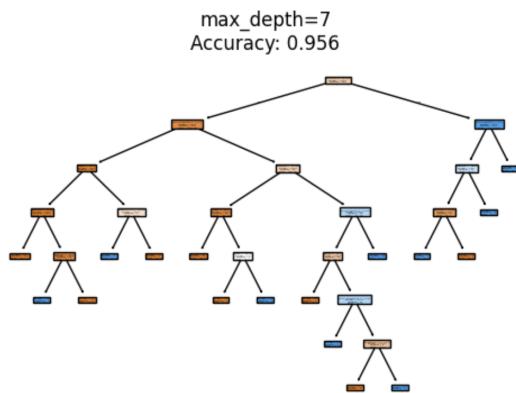
Hình 19: Decision tree với độ sâu là "None" và 2



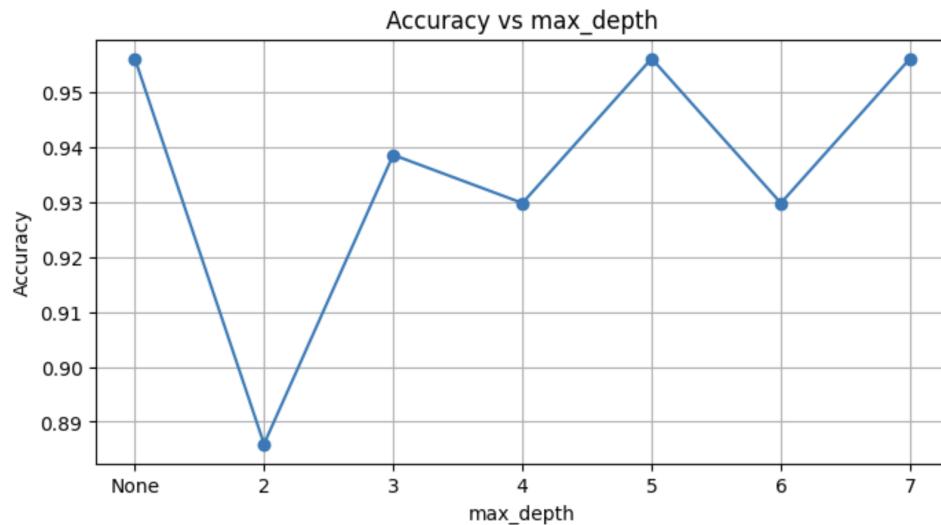
Hình 20: Decision tree với độ sâu là 3 và 4



Hình 21: Decision tree với độ sâu là 5 và 6



Hình 22: Decision tree với độ sâu là 7



Hình 23: Biểu đồ đường thể hiện mối liên hệ của Depth và Accuracy

**Nhận xét:** Dựa trên dữ liệu thống kê và biểu đồ đường phân tích độ chính xác của các Decision tree ở các độ sâu khác nhau, có thể thấy rằng độ sâu của cây quyết định đóng vai trò quan trọng trong hiệu suất mô hình:

- Mô hình với độ sâu tối đa là 2: đạt độ chính xác thấp nhất (88.6%), cho thấy mô hình quá đơn giản không đủ phức tạp.
- Mô hình với độ sâu tối đa là 3: độ chính xác tăng rõ rệt lên 93.9% thể hiện rằng mô hình đã
- Mô hình với độ sâu tối đa là 4 và 6: đạt độ chính xác giảm xuống còn 93%.
- Mô hình với độ sâu từ 5, 7 và độ sâu không giới hạn: độ chính xác đạt mức cao nhất là 95%.

**kết luận:** Ứng với độ sâu nhỏ Decision tree chỉ được tạo ra ít mức phân nhánh điều này làm giảm khả năng mở rộng phân nhánh chi tiết hơn hay nói cách khác mô hình chỉ học một cách tổng quát mà không chi tiết cụ thể bỏ qua những chi tiết quan trọng trong tập train dẫn đến mô hình không thể nhận diện tốt được mối quan hệ giữa các feature nên độ chính xác chưa được cao. Khi độ sâu bắt đầu tăng lên, Decision tree tạo thêm nhiều nhánh hơn, từ đó dữ liệu được mô hình nhận diện chi tiết hơn các mối liên hệ giữa feature và label. Và trong biểu đồ có thể thấy kể từ độ sâu 3 trở đi thì độ chính xác của mô hình luôn ổn định ở mức 93% đến mức cao nhất là 95.6%.

## 4.2 Tập dữ liệu Wine Quality

### 4.2.1 Chuẩn bị dữ liệu

Bộ dữ liệu **UCI Wine Quality Dataset** (chất lượng rượu) chứa các thông tin hóa học và cảm quan (physicochemical & sensory) của rượu vang trắng và rượu vang đỏ đến từ vùng Vinho Verde của Bồ Đào Nha, với mục tiêu là dự đoán chất lượng của rượu thông qua việc đưa ra **điểm chất lượng (quality score)** từ **1-10**.

Tổng cộng có 4898 mẫu trong dataset của rượu vang trắng và 1599 mẫu trong dataset của rượu vang đỏ, mỗi mẫu đều có 12 features, từ đó giúp phân loại mẫu. Có 11 features mô tả về thành phần lý hóa của rượu bao gồm: ***Độ axit cố định, Độ axit bay hơi, Axit citric, Lượng đường dư, Hàm lượng chloride, Lưu huỳnh dioxit tự do, Tổng lượng lưu huỳnh dioxit, Mật độ rượu, Độ pH, Nồng độ sunfat, Nồng độ cồn***; 1 feature còn lại là **điểm chất lượng** của rượu.

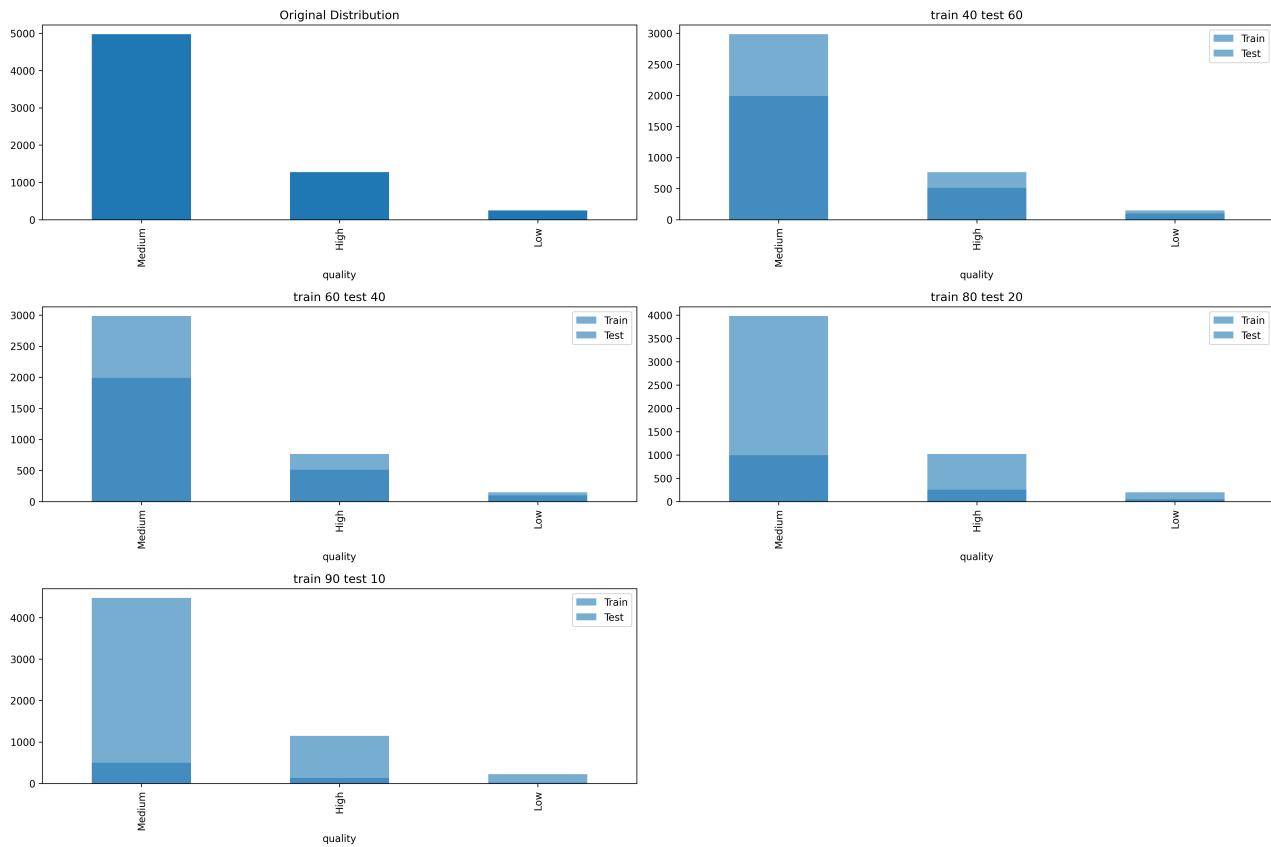
Để so sánh trực quan và tổng quát hơn, nhóm em đã gộp hai dataset trên lại và thêm feature type (ghi nhận mẫu là rượu vang trắng hay đỏ).

Khi bắt đầu xử lí dataset, theo yêu cầu đề bài, ta chia điểm chất lượng thành 3 nhóm **Low (0-4), Medium (5-6) và High (7-10)**. Hàm `classify_quality()` nhận vào điểm và trả ra nhóm phù hợp.

Sau khi xáo trộn Features và Labels, dataset sẽ được chia thành các tập huấn luyện (train) và tập kiểm tra (test) theo các tỷ lệ được yêu cầu. Các tỷ lệ huấn luyện/kiểm tra khác nhau được sử dụng gồm: **40/60, 60/40, 80/20** và **90/10**. Dataset sẽ được tách ra thành các thành phần như sau:

- **feature\_train:** Tập mẫu chứa các features để train.
- **label\_train:** Nhãn tương ứng điểm chất lượng của các mẫu trong tập train.
- **feature\_test:** Tập mẫu chứa các feaatures để test.
- **label\_test:** Nhãn tương ứng điểm chất lượng của các mẫu trong tập test.

Như vậy, sau khi hoàn tất quá trình chia dữ liệu, với mỗi tỷ lệ chúng ta sẽ có 4 tập con, tổng cộng là 16 bộ dữ liệu.



Hình 24: Tỷ lệ giữa tập train và tập test

Sau khi hoàn thành việc chia và chuẩn bị bộ dữ liệu ở các tỷ lệ khác nhau, chúng ta sẽ tiến hành thể hiện (**Visualization**) sự phân bố dữ liệu giữa các lớp (**class**), tập trung vào:

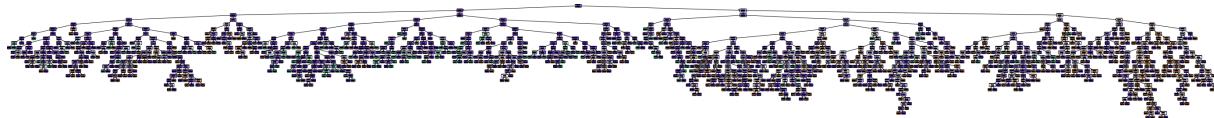
- Tập huấn luyện (training set)
- Tập kiểm tra (testing set)

Thông qua biểu đồ cột chồng thể hiện số lượng mẫu Label "High", "Medium", "Low" trong tập dữ liệu train và test với các mức tỷ lệ khác nhau ta nhận thấy rằng số lượng mẫu Label "Medium" có số lượng lớn hơn rất nhiều so với hai Labels còn lại "Low" và "High" và tỷ lệ chênh lệch duy trì qua các mức tỷ lệ khác nhau của tập train và tập test.

#### 4.2.2 Xây dựng mô hình cây quyết định (Decision Tree)

Dể thể hiện rõ cấu trúc và cách thức hoạt động của mô hình, mỗi cây quyết định sau khi được huấn luyện sẽ được trực quan hóa (**Visualization**) bằng thư viện **graphviz**. Việc trực quan hóa này cho phép ta quan sát trực tiếp cách mô hình đưa ra quyết định dựa trên dữ liệu huấn luyện, từ đó nâng cao khả năng diễn giải và phân tích mô hình.

Cây quyết định được hiển thị dưới dạng sơ đồ phân nhánh, trong đó mỗi nút biểu diễn một điều kiện phân tách dựa trên một features cụ thể, các nhánh thể hiện hướng đi tùy thuộc vào việc điều kiện đó đúng hay sai, và các nút lá thể hiện lớp Label mà mô hình dự đoán. Thông tin hiển thị tại mỗi nút bao gồm tên feature, giá trị threshold, chỉ số entropy, số lượng mẫu và lớp label.



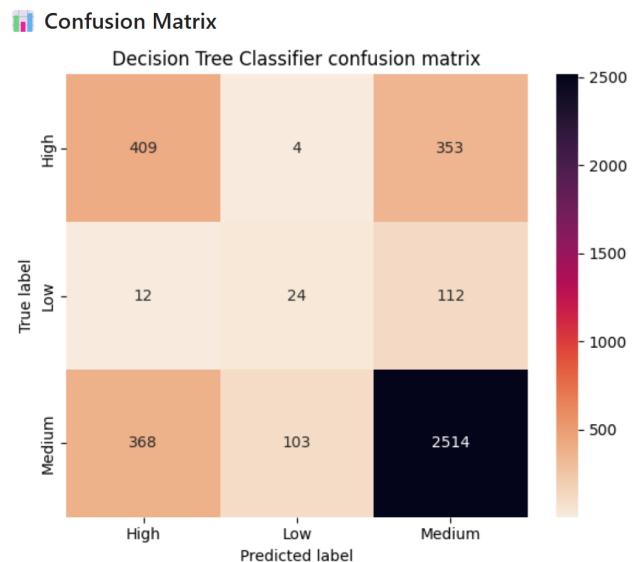
Hình 25: Decision Tree với tỷ lệ 40/60

#### Classification Report

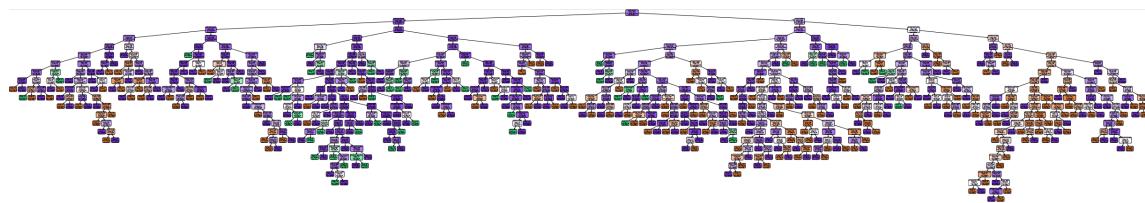
	precision	recall	f1-score	support
High	0.52	0.53	0.53	766
Low	0.18	0.16	0.17	148
Medium	0.84	0.84	0.84	2985
accuracy			0.76	3899
macro avg	0.52	0.51	0.51	3899
weighted avg	0.75	0.76	0.76	3899

Accuracy: 0.76

Hình 26: Phân loại với tỷ lệ 40/60



Hình 27: Ma trận nhầm lẫn với tỷ lệ 40/60



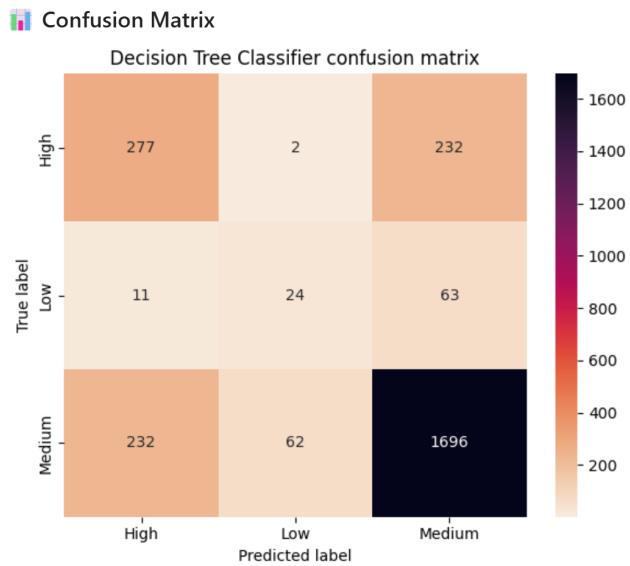
Hình 28: Decision Tree với tỷ lệ 60/40

### Classification Report

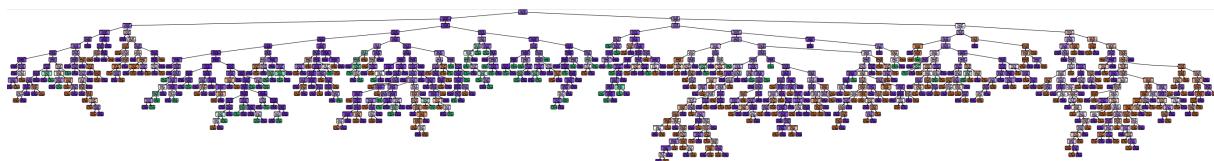
	precision	recall	f1-score	support
High	0.53	0.54	0.54	511
Low	0.27	0.24	0.26	98
Medium	0.85	0.85	0.85	1990
accuracy			0.77	2599
macro avg	0.55	0.55	0.55	2599
weighted avg	0.77	0.77	0.77	2599

Accuracy: 0.77

Hình 29: Phân loại với tỷ lệ 60/40



Hình 30: Ma trận nhầm lẫn với tỷ lệ 60/40



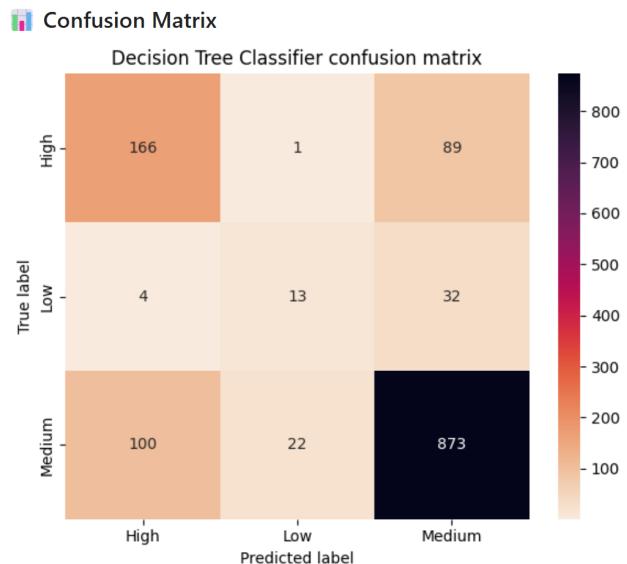
Hình 31: Decision Tree với tỷ lệ 80/20

### Classification Report

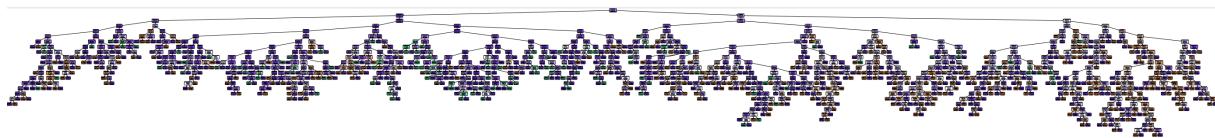
	precision	recall	f1-score	support
High	0.61	0.65	0.63	256
Low	0.36	0.27	0.31	49
Medium	0.88	0.88	0.88	995
accuracy			0.81	1300
macro avg	0.62	0.60	0.60	1300
weighted avg	0.81	0.81	0.81	1300

Accuracy: 0.81

Hình 32: Phân loại với tỷ lệ 80/20



Hình 33: Ma trận nhầm lẫn với tỷ lệ 80/20



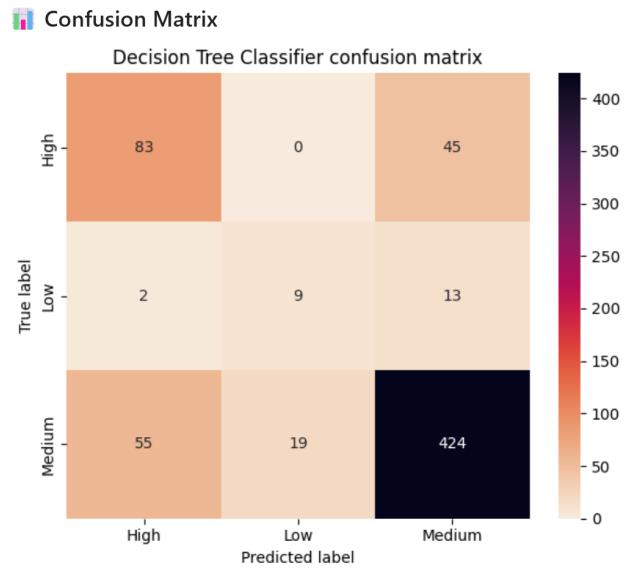
Hình 34: Decision Tree với tỷ lệ 90/10

### Classification Report

	precision	recall	f1-score	support
High	0.59	0.65	0.62	128
Low	0.32	0.38	0.35	24
Medium	0.88	0.85	0.87	498
accuracy			0.79	650
macro avg	0.60	0.62	0.61	650
weighted avg	0.80	0.79	0.80	650

Accuracy: 0.79

Hình 35: Phân loại với tỷ lệ 90/10



Hình 36: Ma trận nhầm lẫn với tỷ lệ 90/10

#### 4.2.3 Đánh giá mô hình Decision Tree

Kết luận mô hình theo tỷ lệ

Mô hình tỷ lệ 40/60

#### Classification Report

	precision	recall	f1-score	support
High	0.52	0.53	0.53	766
Low	0.18	0.16	0.17	148
Medium	0.84	0.84	0.84	2985
accuracy			0.76	3899
macro avg	0.52	0.51	0.51	3899
weighted avg	0.75	0.76	0.76	3899

Accuracy: 0.76

Hình 37: Classification report tỷ lệ 40/60

Độ chính xác của mô hình (Accuracy) là 76% cho thấy khả năng phân loại tạm ổn, nhưng chưa thật sự chính xác. Mô hình vẫn cần được xem xét kỹ để cải thiện thêm – đặc biệt trong những trường hợp mà precision hoặc recall chưa cao.

Trong lớp "High":

- Mô hình đạt precision cao 52% nghĩa là số lượng mẫu bị dự đoán sai thành "High" chiếm khoảng một nửa, con số khá lớn.
- Recall là 53%, nghĩa là mô hình chỉ dự đoán đúng 53% số lượng mẫu thật sự thuộc lớp "High", bỏ sót khoảng một nửa số mẫu thuộc "High".
- F1-Score đạt 53%, cho thấy mô hình có mức độ cân bằng vừa phải giữa precision và recall. Tuy chưa cao, nhưng kết quả này phản ánh mô hình có khả năng phân loại lớp "High" ở mức chấp nhận được, tuy vẫn còn có khả năng tối ưu hơn.

Trong lớp "Low":

- Precision là 18%, nghĩa là phần lớn mẫu khi được dự đoán là "Low" thật ra không thuộc "Low".
- Recall chỉ đạt 16%, mô hình bỏ qua rất nhiều mẫu thật sự thuộc "Low".

- F1-Score đạt 17%, cho thấy mô hình gần như không có khả năng nhận diện chính xác lớp này, với hiệu suất cực kỳ thấp và thiếu ổn định. Điều này phản ánh sự mất cân bằng dữ liệu hoặc đặc trưng lớp “Low” chưa được nhận diện tốt bởi mô hình.

Trong lớp "Medium":

- Precision đạt 84%, nghĩa là phần lớn các mẫu được dự đoán là “Medium” đều đúng với thực tế.
- Recall cũng đạt 84%, cho thấy mô hình nhận diện rất đầy đủ các mẫu thuộc lớp này.
- F1-Score ở mức 84%, cho thấy mô hình có hiệu suất rất ổn định và chính xác với lớp “Medium”, đạt sự cân bằng tốt giữa độ bao phủ và độ chính xác trong dự đoán.

### Mô hình tỷ lệ 60/40

Classification Report				
	precision	recall	f1-score	support
High	0.53	0.54	0.54	511
Low	0.27	0.24	0.26	98
Medium	0.85	0.85	0.85	1990
accuracy			0.77	2599
macro avg	0.55	0.55	0.55	2599
weighted avg	0.77	0.77	0.77	2599
Accuracy: 0.77				

Hình 38: Classification report tỷ lệ 60/40

Dộ chính xác của mô hình (Accuracy) là 77%, cho thấy khả năng phân loại tổng thể ở mức khá tốt. Tuy nhiên, vẫn cần được đánh giá sâu hơn từng lớp – đặc biệt với những lớp có số lượng mẫu ít hoặc kết quả còn thấp.

Trong lớp "High":

- Mô hình đạt precision 53%, nghĩa là gần một nửa số mẫu được dự đoán là “High” thực chất không đúng.
- Recall là 54%, tức là mô hình chỉ phát hiện đúng khoảng một nửa số mẫu thực sự thuộc lớp “High”.

- F1-Score đạt 54%, phản ánh khả năng phân loại lớp “High” ở mức chấp nhận được, nhưng vẫn còn cần cải thiện.

Trong lớp "Low":

- Precision là 27%, nghĩa là hơn 70% số mẫu được gán nhãn “Low” là sai.
- Recall chỉ đạt 24%, mô hình bỏ sót rất nhiều mẫu thực sự thuộc “Low”.
- F1-Score đạt 26%, thể hiện mô hình đang gặp khó khăn nghiêm trọng trong việc nhận diện lớp “Low”.

Trong lớp "Medium":

- Precision đạt 85%, nghĩa là phần lớn các mẫu được dự đoán là “Medium” đều đúng với thực tế.
- Recall cũng đạt 85%, cho thấy mô hình nhận diện rất tốt các mẫu thực sự thuộc lớp này.
- F1-Score ở mức 85%, chứng minh hiệu suất rất ổn định và chính xác với lớp “Medium”.

### Mô hình tỷ lệ 80/20

Classification Report				
	precision	recall	f1-score	support
High	0.61	0.65	0.63	256
Low	0.36	0.27	0.31	49
Medium	0.88	0.88	0.88	995
accuracy			0.81	1300
macro avg	0.62	0.60	0.60	1300
weighted avg	0.81	0.81	0.81	1300
Accuracy: 0.81				

Hình 39: Classification report tỷ lệ 80/20

Dộ chính xác của mô hình (Accuracy) đạt 81%, phản ánh năng lực phân loại tổng thể khá ổn. Tuy nhiên, hiệu quả phân loại giữa các lớp chưa đồng đều, đặc biệt lớp “Low” vẫn còn nhiều hạn chế do số lượng mẫu ít và đặc trưng khó nhận diện.

Trong lớp "High":

- Precision ở mức 61%, cho thấy khoảng 39% mẫu được dự đoán là “High” thực ra không chính xác.
- Recall đạt 65%, mô hình nhận diện được phần lớn các trường hợp thực sự thuộc lớp này.
- F1-Score là 63%, thể hiện mô hình có khả năng phân loại lớp “High” tương đối tốt, nhưng vẫn có thể cải thiện thêm để đạt độ ổn định cao hơn.

Trong lớp "Low":

- Precision chỉ đạt 36%, tức là đa số các mẫu gán nhãn “Low” không đúng với thực tế.
- Recall ở mức 27%, mô hình bỏ sót nhiều trường hợp thực sự thuộc lớp này.
- F1-Score là 31%, phản ánh mô hình đang gặp khó khăn trong việc nhận diện chính xác lớp “Low”.

Trong lớp "Medium":

- Precision đạt 88%, rất cao, nghĩa là phần lớn các dự đoán “Medium” đều chính xác.
- Recall cũng là 88%, chứng minh khả năng bao phủ rất tốt các trường hợp thực tế của lớp này.
- F1-Score cao nhất ở mức 88%, thể hiện mô hình đang hoạt động hiệu quả và đáng tin cậy với lớp “Medium”.

Mô hình tỷ lệ 90/10

Classification Report				
	precision	recall	f1-score	support
High	0.59	0.65	0.62	128
Low	0.32	0.38	0.35	24
Medium	0.88	0.85	0.87	498
accuracy			0.79	650
macro avg	0.60	0.62	0.61	650
weighted avg	0.80	0.79	0.80	650
Accuracy: 0.79				

Hình 40: Classification report tỷ lệ 90/10

Mô hình đạt độ chính xác 79%, cho thấy khả năng dự đoán tổng thể khá ổn định. Tuy nhiên, việc phân biệt giữa các lớp vẫn còn một số điểm yếu, đặc biệt là lớp “Low”.

Trong lớp "High":

- Precision đạt 59%, tức là khoảng 41% các dự đoán “High” là sai.
- Recall là 65%, cho thấy mô hình phát hiện được phần lớn các trường hợp thực sự thuộc lớp này.
- F1-Score ở mức 62%, phản ánh hiệu quả phân loại lớp này ở mức khá nhưng vẫn cần cải thiện.

Trong lớp "Low":

- Precision khá thấp, chỉ 32%, tức là mô hình hay nhầm lẫn khi gán nhãn “Low”.
- Recall là 38%, cho thấy nhiều trường hợp thực sự thuộc lớp “Low” đã bị bỏ sót.
- F1-Score ở mức 35%, minh họa sự yếu kém trong việc nhận diện lớp này.

Trong lớp "Medium":

- Precision đạt 88%, nghĩa là các dự đoán “Medium” gần như luôn đúng.
- Recall ở mức 85%, mô hình cũng không bỏ sót nhiều trường hợp thực sự là “Medium”.
- F1-Score lên đến 87%, cho thấy mô hình nhận diện rất hiệu quả với lớp này.

**Kết luận:** Trong bốn tỉ lệ chia dữ liệu huấn luyện và kiểm tra tương ứng với bốn mô hình cây quyết định khác nhau, mô hình sử dụng tỉ lệ 80/20 là phù hợp nhất. Lý do là vì mô hình này đạt độ chính xác cao (81%), cao hơn những mô hình còn lại. Qua bốn mô hình được đánh giá, có thể rút ra một số xu hướng phân loại như sau:

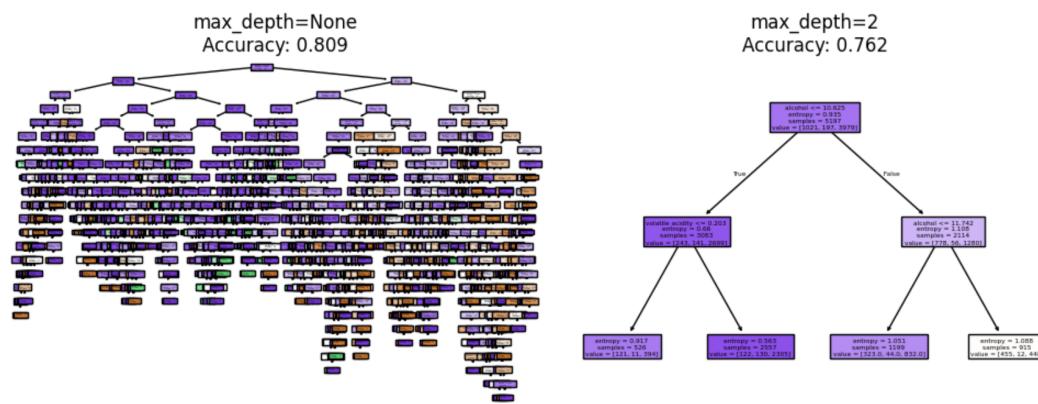
- **Lớp "Medium"** luôn đạt hiệu suất cao nhất, với precision và recall thường ở trong khoảng 85–88%. Điều này cho thấy mô hình học và xử lý rất tốt các features của lớp này. Đạt được điều này là do lớp "Medium" có số lượng mẫu lớn nhất trong tập dữ liệu.
- **Lớp "High"** có độ chính xác trung bình, với F1-score dao động khoảng 62–63%. Mô hình có khả năng phát hiện các mẫu “High” ở mức tương đối ổn, nhưng vẫn còn gặp khó khăn.
- **Lớp "Low"** là lớp có hiệu suất thấp nhất, với precision và recall đều rất thấp (dưới 40%), dẫn đến F1-score thấp. Lý do có thể do:

- Số lượng mẫu thuộc lớp "Low" trong tập huấn luyện nhỏ hơn nhiều so với các lớp khác nên mô hình không học được tốt.
- Features của lớp "Low" chưa đủ khác biệt hoặc bị trùng lặp nhiều với các lớp khác, gây ra dễ nhầm lẫn khi phân loại.
- Nhìn chung, dù độ chính xác tổng thể (*accuracy*) của mô hình đạt mức khá cao (79–81%), nhưng *macro average* cho thấy rằng mô hình vẫn còn thiên lệch, chủ yếu phân biệt tốt các lớp có nhiều dữ liệu hơn.

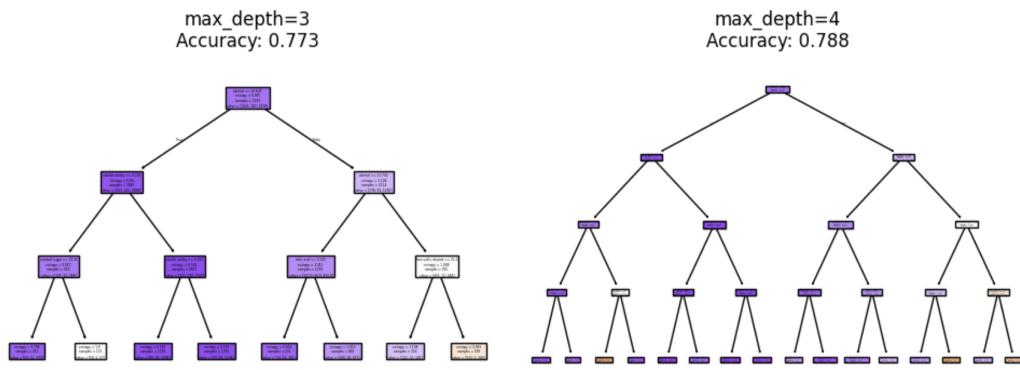
Như vậy, để cải thiện mô hình, có thể cần bổ sung thêm dữ liệu cho lớp "Low", đảm bảo cân bằng số lượng mẫu của mỗi lớp. Đồng thời có thể phân tích lại đặc trưng của từng lớp để tăng khả năng phân loại.

#### 4.2.4 Đánh giá độ chính xác theo độ sâu

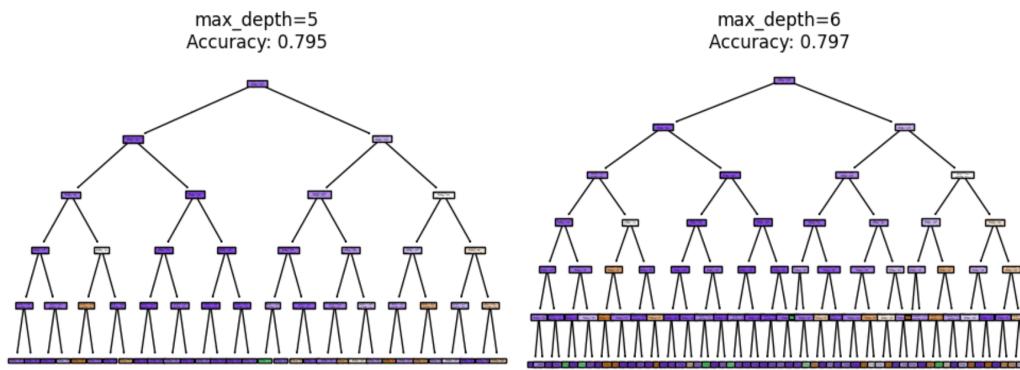
Sử dụng mô hình dự đoán Decision tree với tỷ lệ 80/20 ta sẽ thử nghiệm đổi với mỗi độ sâu (Depth) khác nhau thì nó sẽ ảnh hưởng đến độ chính xác (Accuracy) của mô hình như thế nào. Ta sẽ lần lượt thử nghiệm với các độ sâu: 2, 3, 4, 5, 6 ,7 và không giới hạn độ sâu.



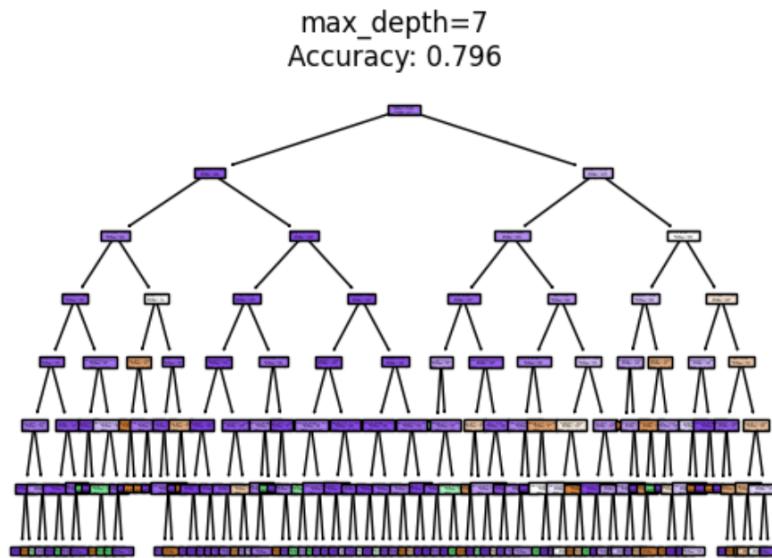
Hình 41: Decision tree với độ sâu là "None" và 2



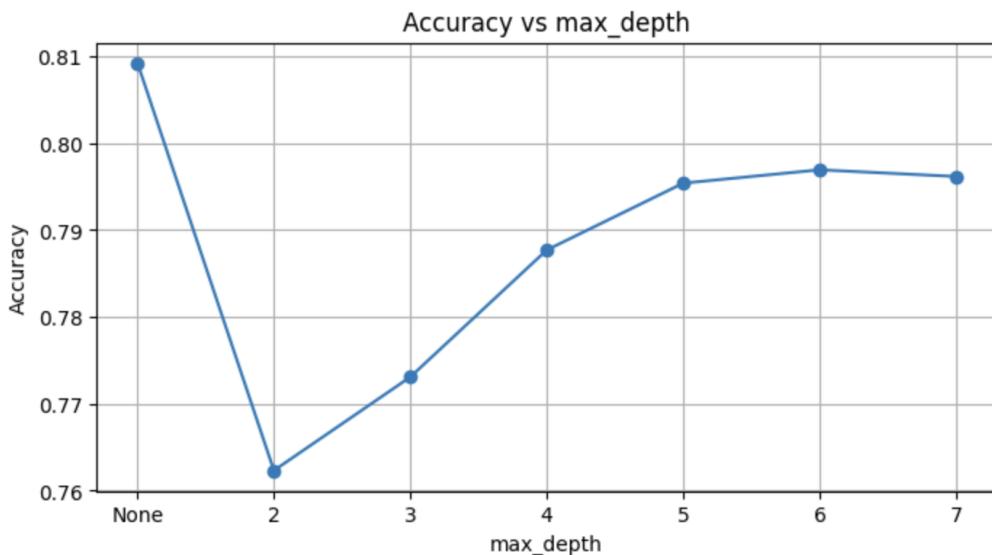
Hình 42: Decision tree với độ sâu là 3 và 4



Hình 43: Decision tree với độ sâu là 5 và 6



Hình 44: Decision tree với độ sâu là 7



Hình 45: Biểu đồ đường thể hiện mối liên hệ của Depth và Accuracy

**Nhận xét:** Dựa trên dữ liệu thống kê và biểu đồ đường phân tích độ chính xác của các Decision tree ở các độ sâu khác nhau, có thể thấy rằng độ sâu của cây quyết định đóng vai trò quan trọng trong hiệu suất mô hình:

- Mô hình với độ sâu tối đa là 2: đạt độ chính xác thấp nhất (76.2%), cho thấy mô hình quá đơn giản không đủ phức tạp.
- Mô hình với độ sâu tối đa là 3: độ chính xác tăng lên 77.3% thể hiện rằng mô hình đã cải thiện.
- Cứ thế các độ sâu 4, 5, 6, độ chính xác tăng dần, lần lượt ứng với 78.8%, 79.5%, 79.7%.
- Mô hình với độ sâu không giới hạn đạt độ chính xác cao nhất với 80.9%.

**Kết luận:** Dựa trên số liệu thống kê, có thể thấy rằng độ sâu của cây quyết định ảnh hưởng rõ rệt đến độ chính xác của mô hình. Với độ sâu nhỏ (như 2 hoặc 3), cây chỉ tạo được ít mức phân nhánh, dẫn đến việc mô hình học một cách quá tổng quát và bỏ qua nhiều chi tiết quan trọng trong dữ liệu huấn luyện — điều này khiến mô hình không thể nhận diện tốt các mối quan hệ giữa các đặc trưng, nên độ chính xác còn thấp (76.2%–77.3%). Khi độ sâu tăng dần, mô hình có khả năng mở rộng phân nhánh, học được các đặc điểm chi tiết hơn trong dữ liệu, từ đó độ chính xác cũng được cải thiện ổn định qua từng mức và đạt cao nhất khi không giới hạn độ sâu.

## 4.3 Tập dữ liệu bổ sung (Age Prediction)

### 4.3.1 Chuẩn bị dữ liệu

Bộ dữ liệu dự đoán tuổi được xây dựng dựa trên dữ liệu từ khảo sát NHANES (National Health and Nutrition Examination Survey), nhằm phục vụ cho mục đích phân loại nhóm tuổi của các cá nhân tại Mỹ. Tập dữ liệu bao gồm 2278 mẫu, trong đó mỗi mẫu được gắn nhãn thuộc một trong hai nhóm tuổi: **Adult** (Người lớn) hoặc **Senior** (Người cao tuổi).

Tổng cộng có 10 đặc trưng được thu thập, bao gồm các giá trị liên quan đến sức khỏe và nhân khẩu học như **RIAGENDR** (giới tính), **PAQ605** (hoạt động thể chất), **BMXBMI** (chỉ số BMI), **LBXGLU** (mức glucose), **DIQ010** (chỉ số tiểu đường), **LBXGLT** (glycohemoglobin), **LBXIN** (insulin), cùng với các cột bổ sung như **SEQN** (số thứ tự), **RIDAGEYR** (tuổi tính bằng năm), và **age\_group** (nhóm tuổi - biến mục tiêu). Trong đó, các đặc trưng như **SEQN** và **RIDAGEYR** sẽ được loại bỏ trong quá trình tiền xử lý để tập trung vào các đặc trưng liên quan trực tiếp đến dự đoán.

Hàm **load\_data()** sẽ được sử dụng để tải dữ liệu và chuẩn bị cho quá trình phân loại nhóm tuổi. Cụ thể, hàm này sẽ tách dữ liệu thành tập đặc trưng (**features**) và nhãn (**labels**), đồng thời xáo trộn dữ liệu theo một thứ tự ngẫu nhiên bằng hàm **shuffle()** để giảm nguy cơ thiên lệch do thứ tự dữ liệu. Sau khi loại bỏ các cột không cần thiết, tập đặc trưng cuối cùng bao gồm 7 đặc trưng: **RIAGENDR**, **PAQ605**, **BMXBMI**, **LBXGLU**, **DIQ010**, **LBXGLT**, và **LBXIN**.

Sau khi xáo trộn dữ liệu, tập dữ liệu sẽ được chia thành tập huấn luyện (**train**) và tập kiểm tra (**test**) theo các tỷ lệ khác nhau. Các tỷ lệ huấn luyện/kiểm tra được sử dụng bao gồm: 40/60, 60/40, 80/20 và 90/10. Quá trình chia dữ liệu sẽ được thực hiện bằng hàm **create\_stratified\_splits()** với các tham số sau:

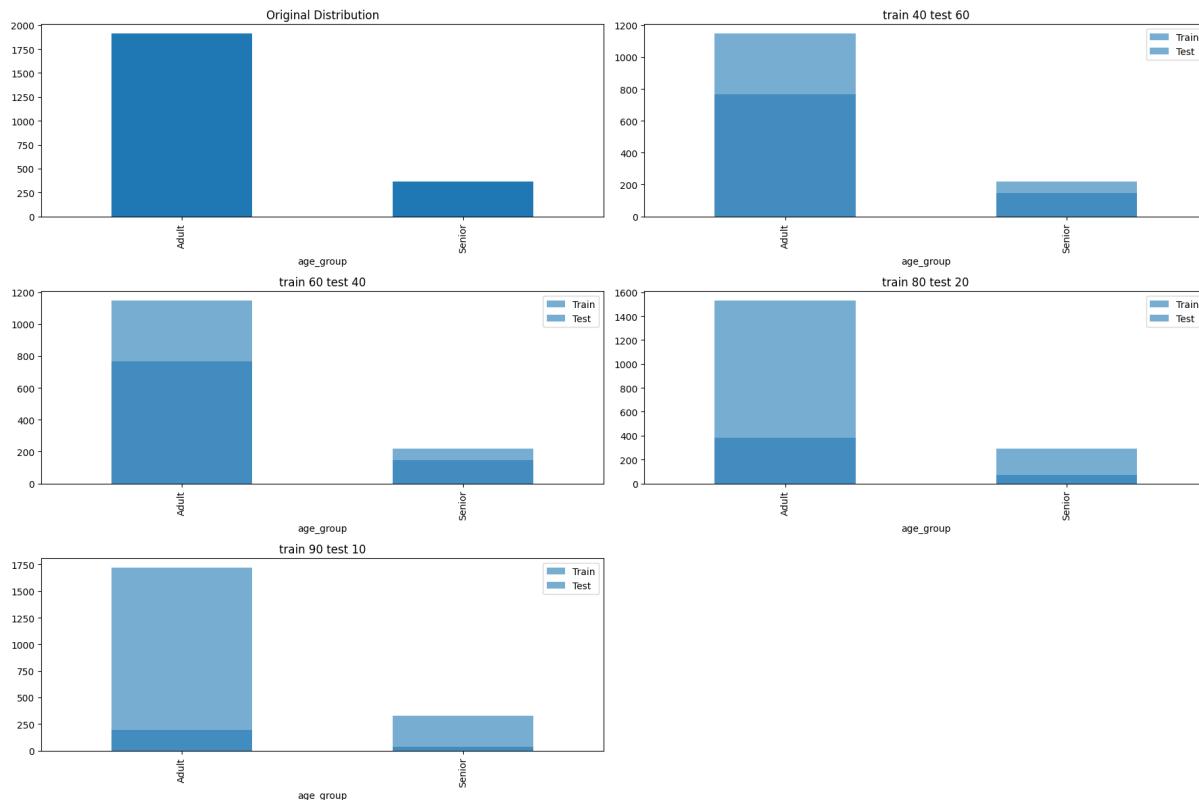
- **stratify**: Đảm bảo giữ nguyên tỷ lệ phân bố giữa hai nhóm tuổi (**Adult** và **Senior**) trong cả tập huấn luyện và tập kiểm tra, nhằm tránh mất cân bằng lớp.
- **random\_state=42**: Đảm bảo kết quả chia dữ liệu có thể lặp lại và không thay đổi qua các lần chạy.

Các bộ dữ liệu sau khi chia sẽ được lưu trữ trong một danh sách có tên là **splits**. Mỗi phần tử trong danh sách này là một bộ 4 thành phần, bao gồm:

- **feature\_train**: Tập đặc trưng huấn luyện (không bao gồm nhãn).
- **feature\_test**: Tập đặc trưng kiểm tra (cùng cấu trúc với **feature\_train**).
- **label\_train**: Nhãn tương ứng với các mẫu trong **feature\_train**.

- **label\_test:** Nhãn tương ứng với các mẫu trong **feature\_test**.

Như vậy, sau khi hoàn tất quá trình chia dữ liệu, với mỗi tỷ lệ chúng ta sẽ có 4 tập con, tổng cộng là 16 bộ dữ liệu.



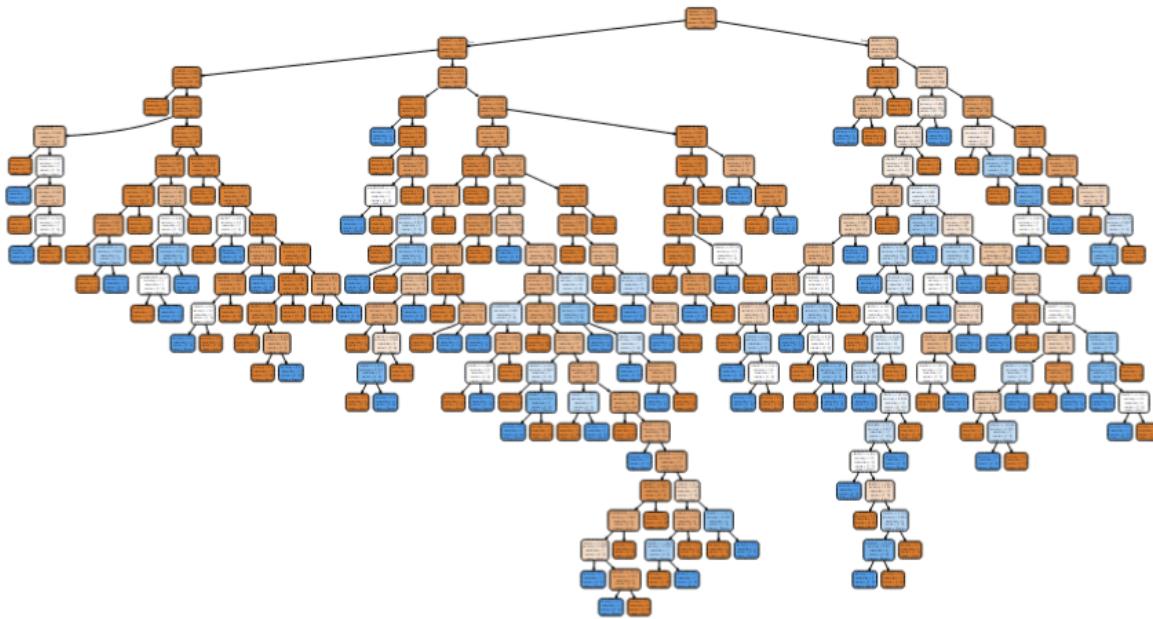
Hình 46: Tỷ lệ giữa tập test và tập train

Biểu đồ "Original Distribution" cho thấy tập dữ liệu có sự mất cân bằng: số lượng mẫu thuộc nhóm Adult lớn hơn đáng kể so với nhóm Senior. Và sự mất cân bằng đó cũng diễn ra đối với các tập con khác. Nhưng đối với các mức tỷ lệ khác nhau thì tỷ lệ giữa các label "Adult" và "Senior" là không đổi.

### 4.3.2 Xây dựng cây quyết định (Decision tree)

Để thể hiện rõ cấu trúc và cách thức hoạt động của mô hình, mỗi cây quyết định sau khi được huấn luyện sẽ được trực quan hóa (**Visualization**) bằng thư viện **graphviz**. Việc trực quan hóa này cho phép ta quan sát trực tiếp cách mô hình đưa ra quyết định dựa trên dữ liệu huấn luyện, từ đó nâng cao khả năng diễn giải và phân tích mô hình.

Cây quyết định được hiển thị dưới dạng sơ đồ phân nhánh, trong đó mỗi nút biểu diễn một điều kiện phân tách dựa trên một feature cụ thể, các nhánh thể hiện hướng đi tùy thuộc vào việc điều kiện đó đúng hay sai, và các nút lá thể hiện lớp Label mà mô hình dự đoán. Thông tin hiển thị tại mỗi nút bao gồm tên feature, giá trị threshold, chỉ số entropy, số lượng mẫu và lớp label.

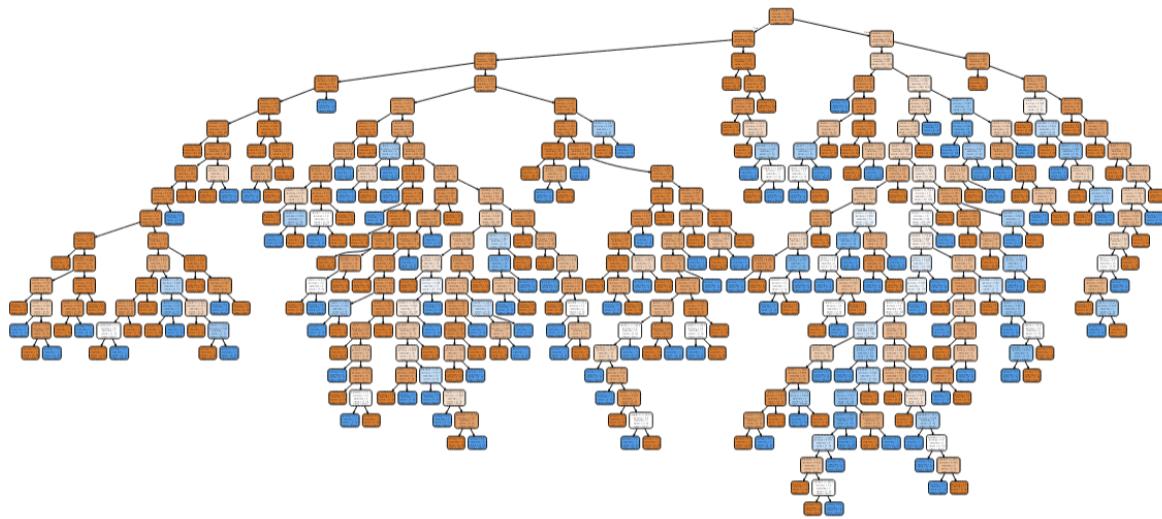


Hình 47: Decision Tree với tỷ lệ 40/60



Hình 48: Phân loại với tỷ lệ 40/60

Hình 49: Ma trận nhầm lẫn với tỷ lệ 40/60



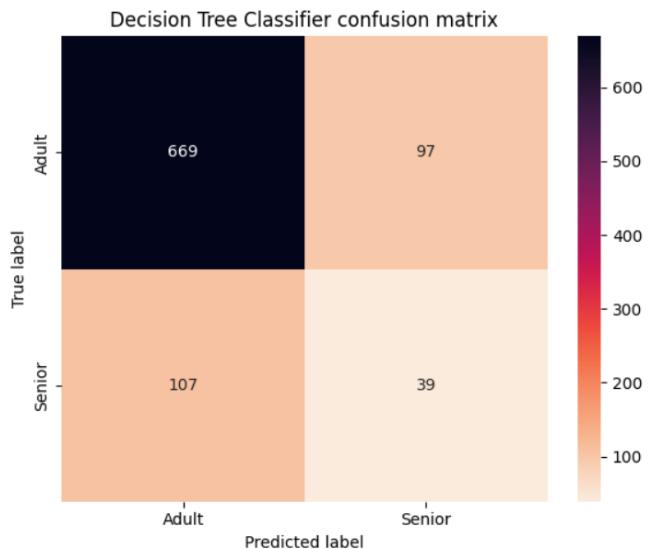
Hình 50: Decision Tree với tỷ lệ 60/40

Classification Report

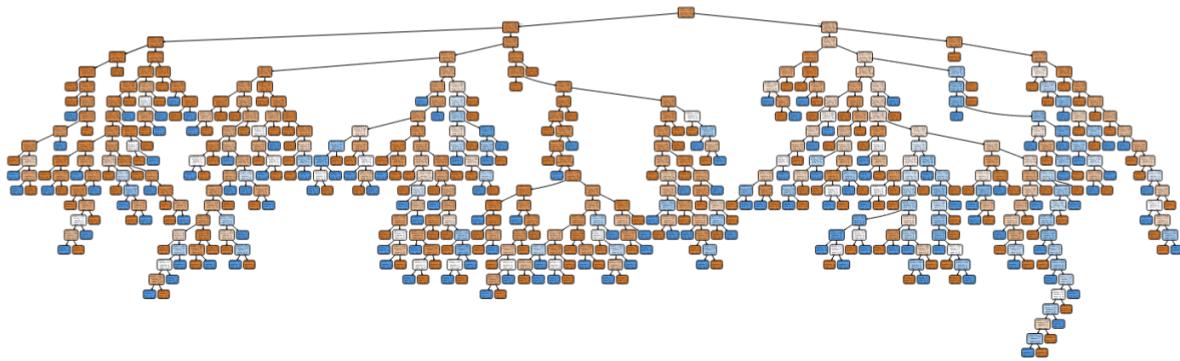
	precision	recall	f1-score	support
Adult	0.86	0.87	0.87	766
Senior	0.29	0.27	0.28	146
accuracy			0.78	912
macro avg	0.57	0.57	0.57	912
weighted avg	0.77	0.78	0.77	912

Accuracy: 0.78

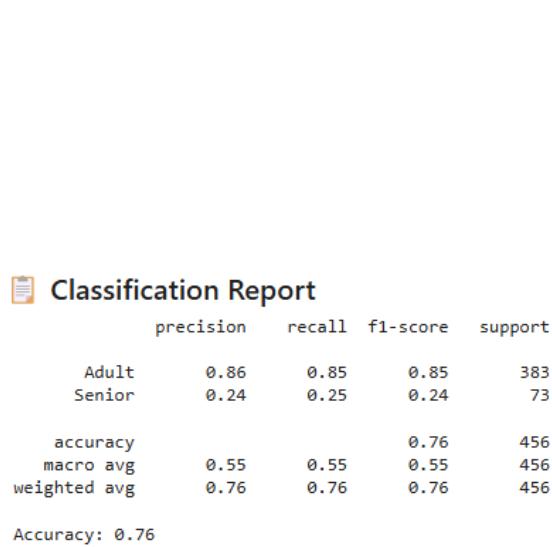
Hình 51: Phân loại với tỷ lệ 60/40



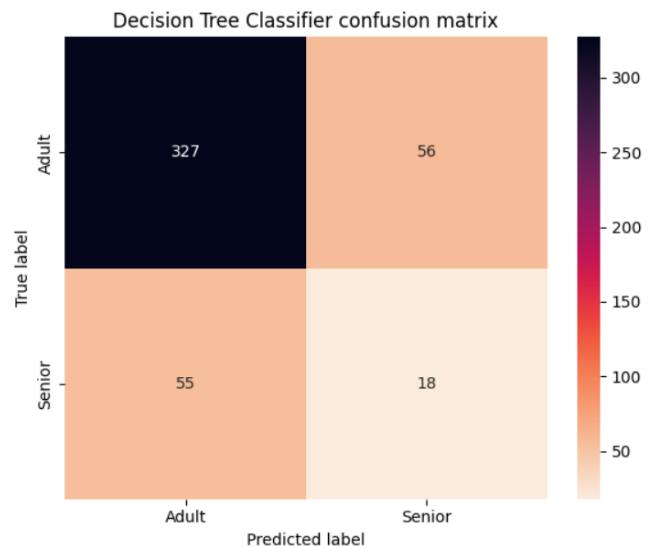
Hình 52: Ma trận nhầm lẫn với tỷ lệ 60/40



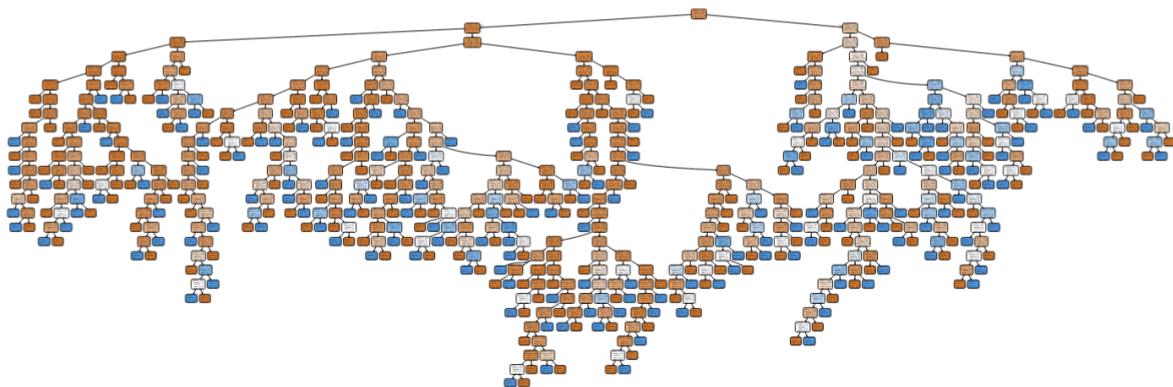
Hình 53: Decision Tree với tỷ lệ 80/20



Hình 54: Phân loại với tỷ lệ 80/20



Hình 55: Ma trận nhầm lẫn với tỷ lệ 80/20



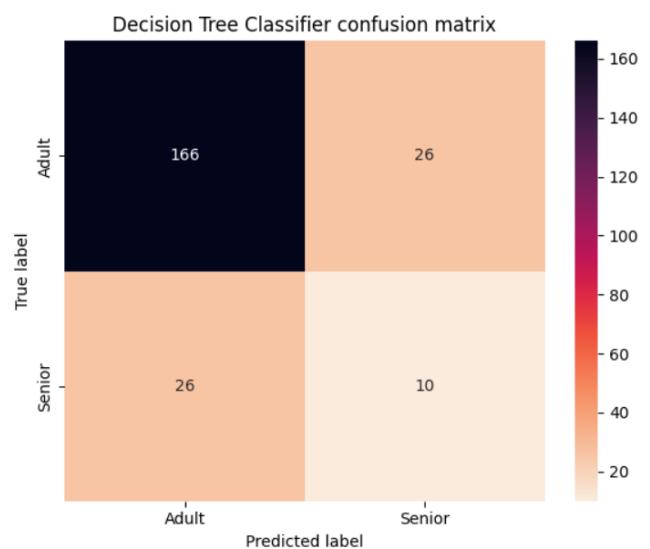
Hình 56: Decision Tree với tỷ lệ 90/10

### Classification Report

	precision	recall	f1-score	support
Adult	0.86	0.86	0.86	192
Senior	0.28	0.28	0.28	36
accuracy			0.77	228
macro avg	0.57	0.57	0.57	228
weighted avg	0.77	0.77	0.77	228

Accuracy: 0.77

Hình 57: Phân loại với tỷ lệ 90/10



Hình 58: Ma trận nhầm lẫn với tỷ lệ 90/10

### 4.3.3 Đánh giá mô hình Decision Tree

Kết luận mô hình theo tỷ lệ

Mô hình tỷ lệ 40/60

Classification Report				
	precision	recall	f1-score	support
Adult	0.87	0.85	0.86	1149
Senior	0.31	0.36	0.33	218
accuracy			0.77	1367
macro avg	0.59	0.60	0.60	1367
weighted avg	0.78	0.77	0.78	1367
Accuracy: 0.77				

Hình 59: Classification report tỷ lệ 40/60

Độ chính xác tổng thể (*Accuracy*) của mô hình là 77%, cho thấy hiệu suất trung bình, nhưng cần cải thiện do sự mất cân bằng giữa hai lớp (**Adult** và **Senior**).

Trong lớp **Adult** (A):

- **Precision** đạt 87%, nghĩa là 87% số mẫu được dự đoán là **Adult** thực sự thuộc lớp này (ít *false positive*).
- **Recall** là 85%, cho thấy mô hình dự đoán đúng 85% số mẫu thực sự thuộc lớp **Adult**, bỏ sót khoảng 15%.
- **F1-Score** đạt 86%, thể hiện sự cân bằng tốt giữa *precision* và *recall*, nhưng vẫn có thể cải thiện do lớp này chiếm ưu thế (1149 mẫu).

Trong lớp **Senior** (S):

- **Precision** chỉ đạt 31%, nghĩa là chỉ 31% số mẫu được dự đoán là **Senior** thực sự thuộc lớp này (nhiều *false positive*).
- **Recall** là 36%, cho thấy mô hình chỉ phát hiện được 36% số mẫu thực sự thuộc lớp **Senior**, bỏ sót tới 64%.
- **F1-Score** đạt 33%, rất thấp, phản ánh hiệu suất kém trên lớp thiểu số này (218 mẫu), do sự mất cân bằng dữ liệu.

Mô hình tỷ lệ 60/40

 Classification Report				
	precision	recall	f1-score	support
Adult	0.86	0.87	0.87	766
Senior	0.29	0.27	0.28	146
accuracy			0.78	912
macro avg	0.57	0.57	0.57	912
weighted avg	0.77	0.78	0.77	912
Accuracy: 0.78				

Hình 60: Classification report tỷ lệ 60/40

Độ chính xác tổng thể (*Accuracy*) của mô hình là 78%, cho thấy hiệu suất khá, nhưng vẫn cần cải thiện do sự mất cân bằng giữa hai lớp (**Adult** và **Senior**).

Trong lớp Adult (A):

- **Precision** đạt 86%, nghĩa là 86% số mẫu được dự đoán là **Adult** thực sự thuộc lớp này (ít *false positive*).
- **Recall** là 87%, cho thấy mô hình dự đoán đúng 87% số mẫu thực sự thuộc lớp **Adult**, bỏ sót khoảng 13%.
- **F1-Score** đạt 87%, thể hiện sự cân bằng tốt giữa *precision* và *recall*, phù hợp với lớp chiếm ưu thế (766 mẫu).

Trong lớp Senior (S):

- **Precision** chỉ đạt 29%, nghĩa là chỉ 29% số mẫu được dự đoán là **Senior** thực sự thuộc lớp này (nhiều *false positive*).
- **Recall** là 27%, cho thấy mô hình chỉ phát hiện được 27% số mẫu thực sự thuộc lớp **Senior**, bỏ sót tới 73%.
- **F1-Score** đạt 28%, rất thấp, phản ánh hiệu suất kém trên lớp thiểu số này (146 mẫu), do sự mất cân bằng dữ liệu.

Mô hình tỷ lệ 80/20

 Classification Report				
	precision	recall	f1-score	support
Adult	0.86	0.85	0.85	383
Senior	0.24	0.25	0.24	73
accuracy			0.76	456
macro avg	0.55	0.55	0.55	456
weighted avg	0.76	0.76	0.76	456
Accuracy: 0.76				

Hình 61: Classification report tỷ lệ 80/20

Độ chính xác tổng thể (*Accuracy*) của mô hình là 76%, cho thấy hiệu suất trung bình, cần cải thiện do sự mất cân bằng giữa hai lớp (*Adult* và *Senior*).

Trong lớp *Adult* (A):

- **Precision** đạt 86%, nghĩa là 86% số mẫu được dự đoán là *Adult* thực sự thuộc lớp này (ít *false positive*).
- **Recall** là 85%, cho thấy mô hình dự đoán đúng 85% số mẫu thực sự thuộc lớp *Adult*, bỏ sót khoảng 15%.
- **F1-Score** đạt 85%, thể hiện sự cân bằng khá giữa *precision* và *recall*, phù hợp với lớp chiếm ưu thế (383 mẫu).

Trong lớp *Senior* (S):

- **Precision** chỉ đạt 24%, nghĩa là chỉ 24% số mẫu được dự đoán là *Senior* thực sự thuộc lớp này (nhiều *false positive*).
- **Recall** là 25%, cho thấy mô hình chỉ phát hiện được 25% số mẫu thực sự thuộc lớp *Senior*, bỏ sót tới 75%.
- **F1-Score** đạt 24%, rất thấp, phản ánh hiệu suất kém trên lớp thiểu số này (73 mẫu), do sự mất cân bằng dữ liệu.

Mô hình tỷ lệ 90/10

 Classification Report				
	precision	recall	f1-score	support
Adult	0.86	0.86	0.86	192
Senior	0.28	0.28	0.28	36
accuracy			0.77	228
macro avg	0.57	0.57	0.57	228
weighted avg	0.77	0.77	0.77	228
Accuracy: 0.77				

Hình 62: Classification report tỷ lệ 90/10

Dộ chính xác tổng thể (*Accuracy*) của mô hình là 77%, cho thấy hiệu suất trung bình, cần cải thiện do sự mất cân bằng giữa hai lớp (*Adult* và *Senior*).

Trong lớp *Adult* (A):

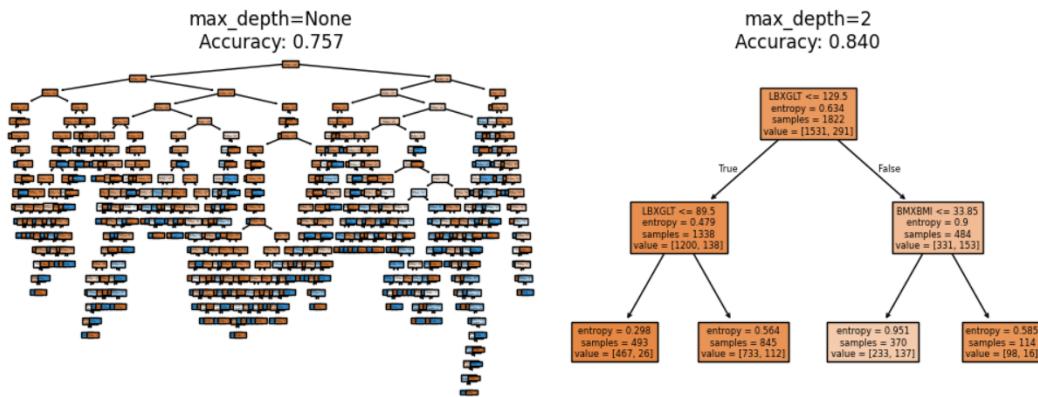
- **Precision** đạt 86%, nghĩa là 86% số mẫu được dự đoán là *Adult* thực sự thuộc lớp này (ít *false positive*).
- **Recall** là 86%, cho thấy mô hình dự đoán đúng 86% số mẫu thực sự thuộc lớp *Adult*, bỏ sót khoảng 14%.
- **F1-Score** đạt 86%, thể hiện sự cân bằng khá giữa *precision* và *recall*, phù hợp với lớp chiếm ưu thế (192 mẫu).

Trong lớp *Senior* (S):

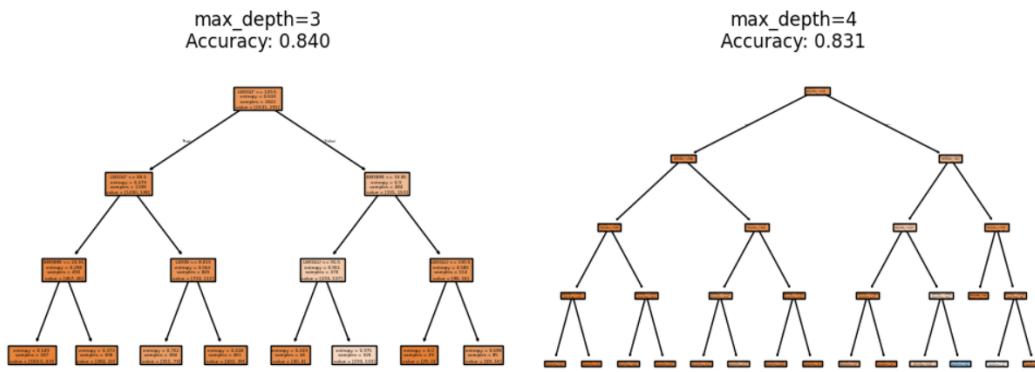
- **Precision** chỉ đạt 28%, nghĩa là chỉ 28% số mẫu được dự đoán là *Senior* thực sự thuộc lớp này (nhiều *false positive*).
- **Recall** là 28%, cho thấy mô hình chỉ phát hiện được 28% số mẫu thực sự thuộc lớp *Senior*, bỏ sót tới 72%.
- **F1-Score** đạt 28%, rất thấp, phản ánh hiệu suất kém trên lớp thiểu số này (36 mẫu), do sự mất cân bằng dữ liệu.

#### 4.3.4 Đánh giá độ chính xác theo độ sâu

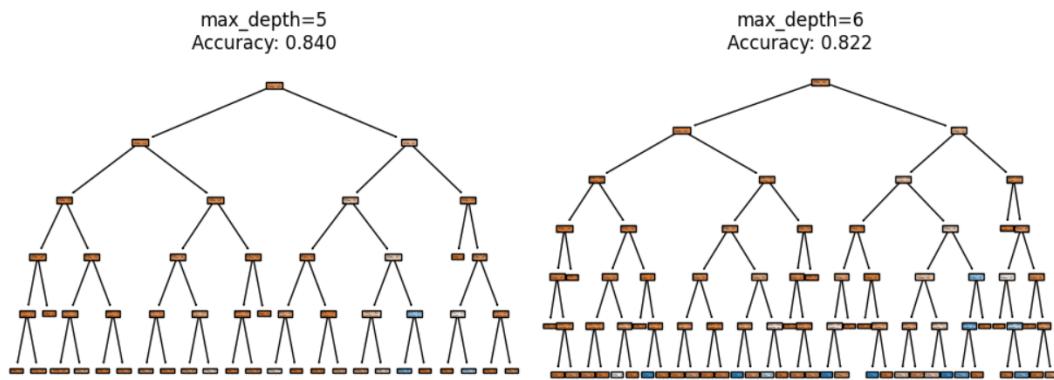
Sử dụng mô hình dự đoán Decision tree với tỷ lệ 80/20 ta sẽ thử nghiệm đổi với mỗi độ sâu (Depth) khác nhau thì nó sẽ ảnh hưởng đến độ chính xác (Accuracy) của mô hình như thế nào. Ta sẽ lần lượt thử nghiệm với các độ sâu: 2, 3, 4, 5, 6, 7 và không giới hạn độ sâu.



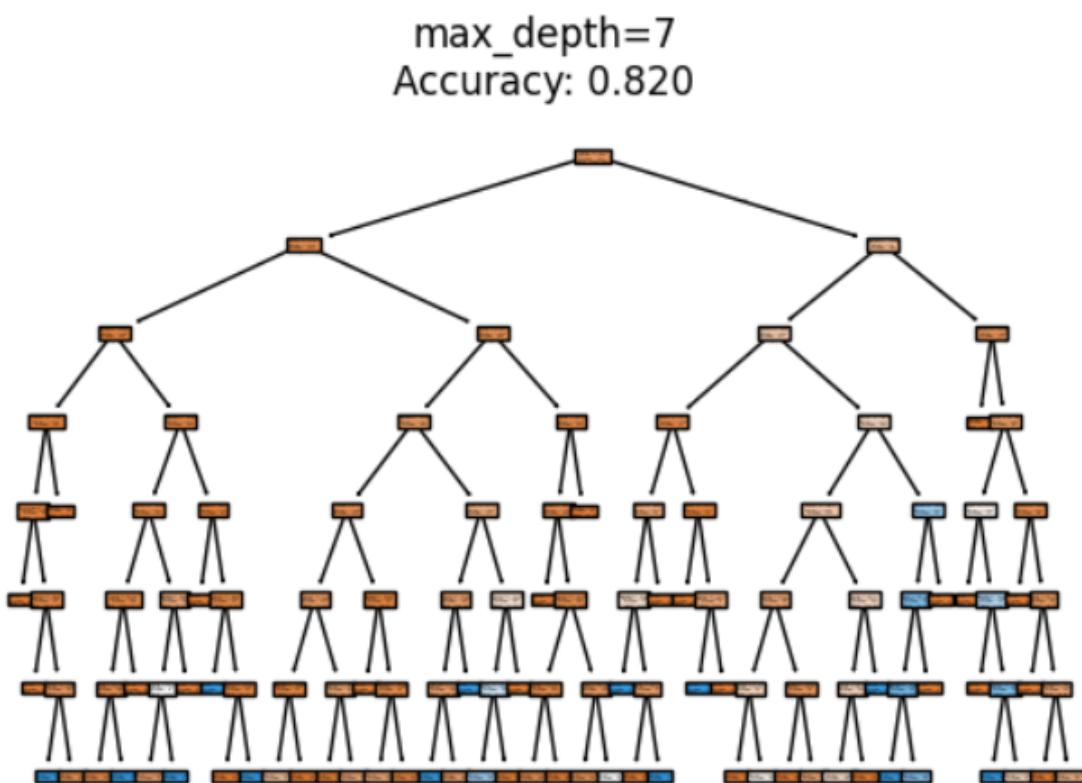
Hình 63: Decision tree với độ sâu là "None" và 2



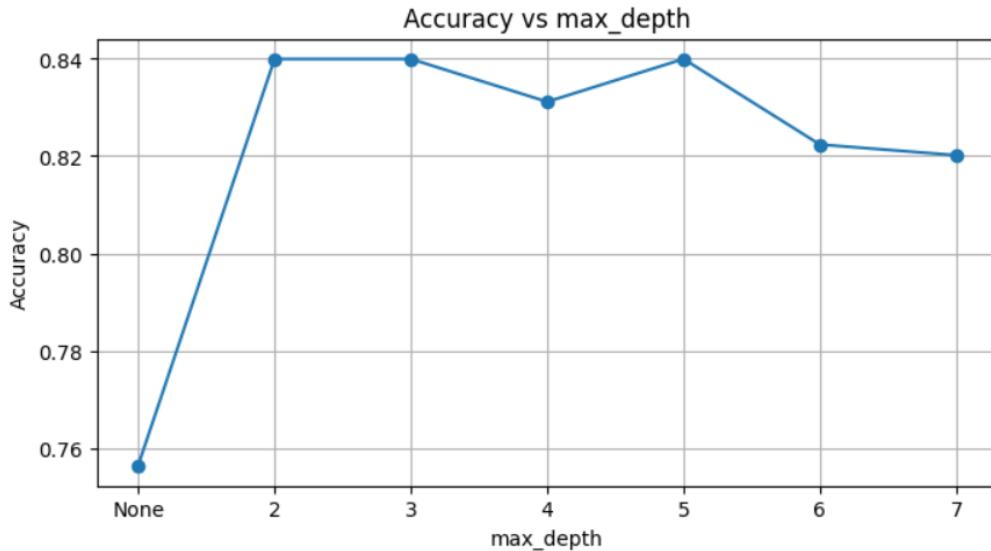
Hình 64: Decision tree với độ sâu là 3 và 4



Hình 65: Decision tree với độ sâu là 5 và 6



Hình 66: Decision tree với độ sâu là 7



Hình 67: Biểu đồ đường thể hiện mối liên hệ của Depth và Accuracy

**Nhận xét:** Dựa trên dữ liệu thống kê và biểu đồ đường phân tích độ chính xác của các Decision tree ở các độ sâu khác nhau, có thể thấy rằng độ sâu của cây quyết định đóng vai trò quan trọng trong hiệu suất mô hình:

- Mô hình với độ sâu tối đa là 2 và 3: đạt độ chính xác 84%
- Mô hình với độ sâu tối đa là 4: Độ chính xác giảm xuống 83.1% Giảm không đáng kể.
- Mô hình với độ sâu tối đa là 5: Quay trở về mốc 84%.
- Mô hình với độ sâu từ 5, 7 và độ sâu không giới hạn: Có dấu hiệu giảm xuống.

**Kết luận:** Kết quả cho thấy độ sâu tối ưu của cây quyết định nằm trong khoảng 3-5. Ở depth là 3 và 5, mô hình đạt accuracy cao nhất (0.84), cho thấy khả năng nắm bắt tốt các mẫu hình trong dữ liệu mà không quá phức tạp. Khi depth là 2, accuracy chỉ đạt 0.757 do cây quá đơn giản để học các đặc điểm quan trọng mô hình bỏ qua nhiều chi tiết trong tập huấn luyện. Ngược lại, khi depth từ 6 trở đi, accuracy bắt đầu giảm do cây bắt đầu học cả các chi tiết không cần thiết điều này dẫn đến mô hình học những feature gây nhiều cho quá trình dự đoán khiến cho độ chính xác không cao.

## 4.4 So sánh giữa các tập dữ liệu

### 4.4.1 Tổng quan

	Breast Cancer	Wine Quality	Age Prediction
Số lượng mẫu	569	4,898	2,278
Label	2 (B/M)	3 (L/M/H)	2 (A/S)
Feature	30	12	7
Cân bằng dữ liệu	Tương đối	Không	Không
Tỷ lệ lớp chiếm ưu thế	~63% (B)	~85% (M)	~84% (A)

Bảng 3: Bảng thống kê của các tập dữ liệu

### 4.4.2 Độ chính xác tổng thể (Accuracy)

Tỷ lệ	Breast Cancer	Wine Quality	Age Prediction
40/60	91%	76%	77%
60/40	94%	77%	78%
80/20	96%	81%	76%
90/10	95%	79%	77%

Bảng 4: Độ chính xác theo các tỷ lệ train/test

Nhận xét:

- Tập Breast Cancer đạt độ chính xác cao nhất (91-96%), do:
  - Dữ liệu cân bằng hơn (tỷ lệ 63/37)
  - Các đặc trưng phân biệt rõ ràng giữa 2 lớp
- Tập Wine Quality và Age Prediction có độ chính xác thấp hơn do:
  - Mất cân bằng dữ liệu nghiêm trọng
  - Các đặc trưng ít phân biệt hơn, đặc biệt với lớp thiểu số

#### 4.4.3 Hiệu suất theo từng lớp (F1-score)

- Breast Cancer:

- Lớp B (lành tính): 93-97%
- Lớp M (ác tính): 88-94%
- Chênh lệch: 5-9%

- Wine Quality:

- Lớp M (trung bình): 84-88%
- Lớp H (cao): 52-63%
- Lớp L (thấp): 17-35%
- Chênh lệch lớn (tới 71%)

- Age Prediction:

- Lớp A (người lớn): 85-87%
- Lớp S (cao tuổi): 24-33%
- Chênh lệch lớn (tới 63%)

#### 4.4.4 Ảnh hưởng của độ sâu cây

Độ sâu	Breast Cancer	Wine Quality	Age Prediction
2	88.6%	76.2%	84%
3	93.9%	77.3%	84%
4	93.0%	78.8%	83.1%
5	95.6%	79.5%	84%
6	93.0%	79.7%	82.2%
7	95.6%	79.6%	82%
None	95.6%	80.9%	75.7

Bảng 5: Độ chính xác theo độ sâu cây

Nhận xét:

- Tập Breast Cancer đạt hiệu suất cao ngay từ độ sâu 3
- Tập Wine Quality cần độ sâu lớn hơn ( $\geq 5$ ) để đạt hiệu suất tốt
- Độ sâu tối ưu khác nhau tùy tập dữ liệu:

- Breast Cancer: 3-5
- Wine Quality:  $\geq 5$
- Age Prediction: tương đối ổn định

## 4.5 Kết luận và đề xuất

- Tập Breast Cancer cho kết quả tốt nhất do:
  - Dữ liệu cân bằng
  - Đặc trưng phân biệt rõ ràng
  - Đạt  $>90\%$  accuracy với mọi tỷ lệ train/test
- Tập Wine Quality và Age Prediction gặp khó khăn với lớp thiểu số do:
  - Mất cân bằng dữ liệu nghiêm trọng
  - Đặc trưng ít phân biệt hơn

**Kết luận:** Kết quả cho thấy hiệu suất mô hình phụ thuộc lớn vào chất lượng và sự tỷ lệ giữa các label, nếu trong tập dữ liệu có quá ít feature điều này dẫn đến mô hình thiếu thông tin để phân biệt các label, tuy nhiên giả sử số lượng feature đủ nhiều nhưng chất lượng kém chất lượng, ít liên quan đến label thì cũng sẽ dẫn đến giảm độ chính xác của mô hình. Qua đó có thể thấy được để một mô hình có được tỷ lệ dự đoán chính xác cao thì cần có sự cân bằng số lượng giữa các label đồng thời số lượng feature đủ nhiều và chất lượng để mô hình có thể phân biệt và đưa ra dự đoán chính xác cho dữ liệu.

## 5 Tham khảo

- UC Irvine Machine Learning Repository
- scikit-learn
- Graphviz source