

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY



CƠ SỞ TRÍ TUỆ NHÂN TẠO

BÁO CÁO ĐỒ ÁN: *DECISION TREE*

Giảng viên hướng dẫn:	Nguyễn Ngọc Đức
Thành viên:	Trần Hùng Anh 22120016
	Trương Tiến Anh 22120017
	Trần Thanh Bình 22120032
	Lê Văn Thành Đạt 22120054

Thành phố Hồ Chí Minh, 12/2024

Mục lục

I	Giới thiệu nhóm và tiến độ công việc	2
II	Bảng đánh giá chi tiết	2
III	Báo cáo bài tập	3
1	Chuẩn bị datasets	3
1.1	Các bộ dữ liệu	3
1.2	Chia tập dữ liệu (40/60, 60/40, 80/20, 90/10)	5
1.3	Trực quan hóa phân phối các lớp	5
2	Xây dựng mô hình cây quyết định	8
2.1	Xây dựng và huấn luyện mô hình với các tỷ lệ khác nhau	8
2.2	Trực quan hóa cây quyết định	8
3	Đánh giá mô hình cây quyết định	11
3.1	Classification report và Confusion matrix	11
3.1.1	Classification report	11
3.1.2	Confusion matrix	12
3.2	Nhận xét các biểu đồ trên	13
3.2.1	Breast Cancer	13
3.2.2	Wine Quality	16
3.2.3	Heart Disease	18
4	Độ sâu và độ chính xác	21
4.1	Cây quyết định với các độ sâu (max_depth) khác nhau	21
4.1.1	Breast Cancer	21
4.1.2	Wine Quality	23
4.1.3	Heart Disease	24
4.2	So sánh và nhận xét ảnh hưởng độ sâu	25
4.2.1	Breast Cancer	25
4.2.2	Wine Quality	25
4.2.3	Heart Disease	26
5	Phân tích và so sánh các bộ dữ liệu	27
5.1	Phân tích đặc điểm từng bộ dữ liệu	27
5.2	Phân tích các yếu tố ảnh hưởng đến hiệu suất cây quyết định trên 3 bộ dữ liệu	28
6	Tổng kết	30



Phân I

Giới thiệu nhóm và tiến độ công việc

Thông tin các thành viên:

STT	MSSV	Họ và tên	Email
1	22120016	Trần Hùng Anh	22120016@student.hcmus.edu.vn
2	22120017	Trương Tiến Anh	22120017@student.hcmus.edu.vn
3	22120032	Trần Thanh Bình	22120032@student.hcmus.edu.vn
4	22120054	Lê Văn Thành Đạt	22120054@student.hcmus.edu.vn

Bảng 1: Thông tin thành viên

Phân công công việc:

STT	Họ và tên	Công việc	Đánh giá
1	Trần Hùng Anh	Chuẩn bị datasets, viết báo cáo	100%
2	Trương Tiến Anh	Xây dựng bộ phân lớp decision tree	100%
3	Trần Thanh Bình	Đánh giá bộ phân lớp decision tree	100%
4	Lê Văn Thành Đạt	Độ sâu và độ chính xác của decision tree	100%

Bảng 2: Phân công công việc

Phân II

Bảng đánh giá chi tiết

STT	Chi tiết	Mức độ hoàn thành
1	Chuẩn bị datasets Chuẩn bị subset Visualize các phân phối	100%
2	Xây dựng bộ phân lớp decision tree	100%
3	Đánh giá bộ phân lớp decision tree Báo cáo phân lớp và confusion matrix Comment	100%
4	Độ sâu và độ chính xác của decision tree Cây, bảng và biểu đồ Comment	100%

Bảng 3: Bảng đánh giá chi tiết



Phân III

Báo cáo bài tập

1 Chuẩn bị datasets

1.1 Các bộ dữ liệu

(a) Tổng quan về 3 bộ dữ liệu

Tên bộ dữ liệu	Đường dẫn	Số lượng mẫu	Số thuộc tính	Mục tiêu
Breast Cancer Dataset	UCI Machine Learning Repository	569	30	Phân loại khối u
Wine Quality Dataset	Wine Quality - UCI Machine Learning Repository	4898	11	Phân loại chất lượng rượu
Heart Disease Dataset	Heart Disease - UCI Machine Learning Repository	303	13	Dự đoán nguy cơ mắc bệnh tim

Bảng 4: Bộ dữ liệu

(b) Chi tiết từng bộ dữ liệu

Breast Cancer Dataset

Bộ dữ liệu Breast Cancer Wisconsin (Diagnostic) là một tập dữ thống kê về căn bệnh Ung thư vú, được để phân loại thành các khối u thành lành tính (B - Benign) hoặc ác tính (M - Malignant). Gồm 569 mẫu với:

- 357 mẫu lành tính (B).
- 212 mẫu ác tính (M)

Tổng cộng có 30 đặc trưng số học (numeric features).

Tên thuộc tính	Ý nghĩa	Loại đặc trưng
radius	Bán kính của khối u (tính từ tâm đến biên dạng trung bình)	Số liên tục
texture	Dộ nhẵn bề mặt của khối u	Số liên tục
perimeter	Chu vi của khối u	Số liên tục
area	Diện tích của khối u	Số liên tục
smoothness	Dộ mượt (sự thay đổi cục bộ của độ dài bán kính)	Số liên tục
compactness	Dộ nén (tỷ lệ giữa diện tích và chu vi bình phương)	Số liên tục
concavity	Dộ lõm (mức độ lõm của hình dạng khối u)	Số liên tục
concave points	Số điểm lõm trên chu vi	Số liên tục
symmetry	Dộ đối xứng của khối u	Số liên tục
fractal dimension	Kích thước fractal (đo lường sự phức tạp của biên dạng khối u)	Số liên tục

Wine Quality Dataset



Đây là bộ dữ liệu được sử dụng để phân loại các mức chất lượng của rượu vang dựa trên các chỉ số về hóa học, bao gồm: độ axit, độ ngọt, lượng cồn, ... Bộ dữ liệu có thang đo chất lượng rượu từ 0-10 (số điểm chất lượng).

Bộ dữ liệu được chia thành ba nhóm dựa trên mức chất lượng:

- Low quality (0-4)
- Standard quality (5-6)
- High quality (đạt 7-10)

Tên đặc trưng	Ý nghĩa	Loại đặc trưng
fixed acidity	Lượng axit cố định trong rượu (chủ yếu axit tartaric).	Số liên tục
volatile acidity	Lượng axit bay hơi (chủ yếu axit acetic), ảnh hưởng đến mùi của rượu.	Số liên tục
citric acid	Lượng axit citric, góp phần tạo độ tươi.	Số liên tục
residual sugar	Lượng đường còn sót lại sau khi quá trình lên men kết thúc.	Số liên tục
chlorides	Hàm lượng muối trong rượu.	Số liên tục
free sulfur dioxide	Lượng SO ₂ tự do trong rượu, giúp ngăn ngừa quá trình oxy hóa và vi khuẩn.	Số liên tục
total sulfur dioxide	Tổng lượng SO ₂ , bao gồm cả SO ₂ tự do và liên kết.	Số liên tục
density	Mật độ của rượu, ảnh hưởng bởi lượng đường và cồn.	Số liên tục
pH	Độ axit của rượu, ảnh hưởng đến hương vị và bảo quản.	Số liên tục
sulphates	Hàm lượng sulfat (muối), góp phần cải thiện mùi vị của rượu.	Số liên tục
alcohol	Hàm lượng cồn, là yếu tố chính ảnh hưởng đến thể tích rượu trong rượu vang.	Số liên tục

Bảng 5: Bảng các thuộc tính và ý nghĩa trong Wine Quality Dataset

Heart Disease Dataset

Đây là bộ dữ liệu được sử dụng để dự đoán nguy cơ phát bệnh tim dựa trên các chỉ số sức khỏe. Bộ dữ liệu gồm 300 mẫu và 14 đặc trưng.

Tên đặc trưng	Ý nghĩa	Loại đặc trưng
age	Tuổi của bệnh nhân (tính bằng năm).	Số liên tục
sex	Giới tính của bệnh nhân (1 = nam; 0 = nữ).	Số phân loại
cp	Loại đau ngực: 1 = đau thắt ngực điển hình; 2 = không điển hình; 3 = không do tim; 4 = không có triệu chứng.	Số phân loại
trestbps	Huyết áp tâm thu khi nghỉ ngơi (mm Hg).	Số liên tục
chol	Mức cholesterol trong máu (mg/dl).	Số liên tục
fbs	Dường huyết lúc đói > 120 mg/dl (1 = đúng; 0 = sai).	Số phân loại



restecg	Kết quả điện tâm đồ nghỉ: 0 = bình thường; 1 = bất thường; 2 = phì đại thất trái.	Số phân loại
thalach	Nhip tim tối đa đạt được.	Số liên tục
exang	Dau thắt ngực do vận động (1 = có; 0 = không).	Số phân loại
oldpeak	Độ chênh ST so với nghỉ ngơi.	Số liên tục
slope	Độ dốc của đoạn ST: 1 = dốc lên; 2 = bằng phẳng; 3 = dốc xuống.	Số phân loại
ca	Số lượng mạch chính được nhuộm màu bằng fluoroscopy (0-3).	Số phân loại
thal	Tình trạng thalassemia: 3 = bình thường; 6 = khiếm khuyết cố định; 7 = có thể đảo ngược.	Số phân loại
num	Chẩn đoán bệnh tim (giá trị từ 0 đến 2).	Số phân loại

Bảng 6: Bảng các thuộc tính và ý nghĩa trong Heart Disease Dataset

1.2 Chia tập dữ liệu (40/60, 60/40, 80/20, 90/10)

Chúng em sẽ tải tập dữ liệu bằng cách sử dụng hàm `load_data()` để chuẩn bị cho việc chia thành các tập train và test. Để đảm bảo rằng các tập huấn luyện và thử nghiệm phản ánh chính xác các đặc điểm của tập dữ liệu gốc, các tập dữ liệu tính năng và nhãn sẽ được xáo trộn theo cùng một thứ tự bằng cách sử dụng hàm `shuffle()` để duy trì mối tương quan giữa dữ liệu tính năng và nhãn của nó.

Sau khi xáo trộn dữ liệu ta sẽ chia bộ dữ liệu ra thành tập train và test theo các tỷ lệ đã yêu cầu. Các tỷ lệ Train/Test khác nhau là (40/60, 60/40, 80/20 và 90/10). Sau khi tập dữ liệu được chia thành các tập huấn luyện và kiểm tra bằng hàm `train_test_split()` với các tham số "`stratify`" để đảm bảo phân phối nhãn được bảo toàn và "`random_state`" được đặt thành 42 để đảm bảo tính nhất quán giữa các lần chạy. Mỗi mục sẽ chứa các thành phần sau:

- **feature_train:** Dữ liệu đặc trưng dùng cho quá trình huấn luyện (không bao gồm thuộc tính target)
- **feature_test:** Dữ liệu đặc trưng dùng cho quá trình kiểm tra (target)
- **label_train:** Nhãn (label) tương ứng với dữ liệu huấn luyện.
- **label_test:** Nhãn (label) tương ứng với dữ liệu kiểm tra.

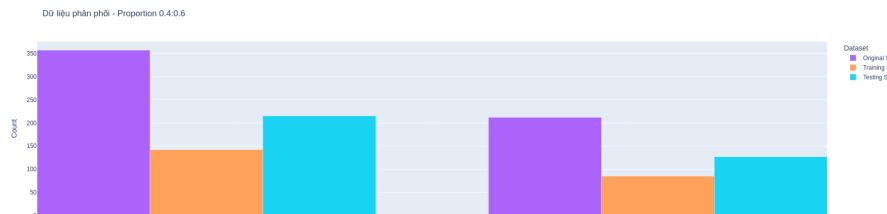
Do đó, sau khi hoàn thành quá trình chia dữ liệu, chúng em sẽ có tổng cộng 16 bộ dữ liệu (4 tỷ lệ chia, mỗi tỷ lệ sẽ tạo ra 4 bộ dữ liệu) : `feature_train`, `feature_test`, `label_train`, và `label_test`. Tất cả các tỷ lệ chia (40/60, 60/40, 80/20, và 90/10) sẽ được áp dụng lần lượt để đảm bảo tính toàn diện trong việc đánh giá mô hình.

1.3 Trực quan hóa phân phối các lớp

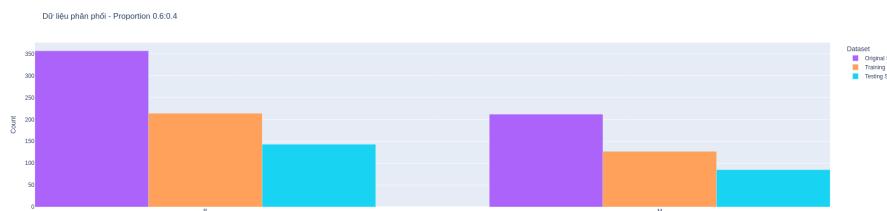
Sau khi hoàn tất việc chia và chuẩn bị tập dữ liệu ở các tỷ lệ khác nhau, chúng em sẽ thực hiện trực quan hóa để phân tích sự phân bố dữ liệu giữa các lớp, tập trung đặc biệt vào tập huấn luyện và tập kiểm tra. Phương pháp sử dụng là biểu đồ cột cô nhầm so sánh số lượng của từng lớp trong các tập dữ liệu: dữ liệu gốc, tập huấn luyện, và tập kiểm tra, ứng với các tỷ lệ chia khác nhau. Việc trực quan hóa này hỗ trợ phân tích sự phân bố của các lớp, đồng thời đánh giá mức độ cân bằng trong quá trình chia dữ liệu, đảm bảo dữ liệu đã được xử lý một cách hợp lý.

Các biểu đồ cột minh họa rõ ràng tỷ lệ từng lớp trong dữ liệu gốc, tập huấn luyện, và tập kiểm tra. Qua đó, chúng em có thể kiểm tra xem tỷ lệ giữa các lớp trong từng tập có được duy trì hay không, đặc biệt với các tỷ lệ chia phổ biến như **40/60, 60/40, 80/20, và 90/10**.

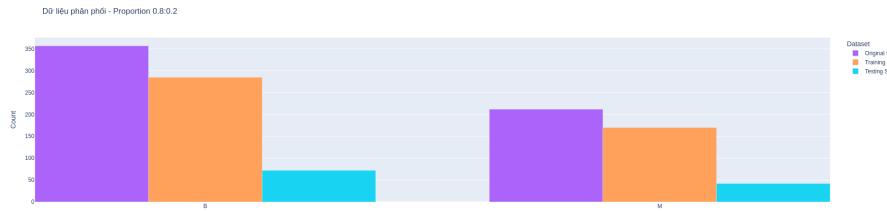
- **Breast Cancer Dataset**



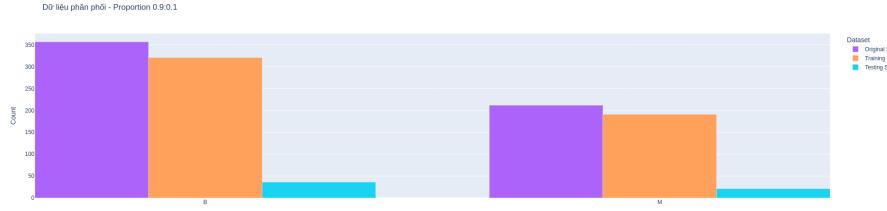
Hình 1: Phân phối dữ liệu 40:60



Hình 2: Phân phối dữ liệu 60:40

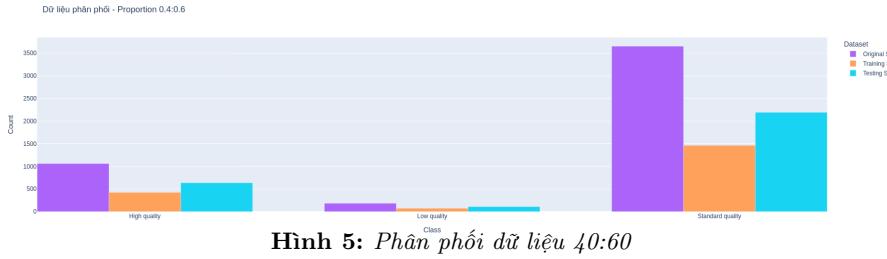


Hình 3: Phân phối dữ liệu 80:20

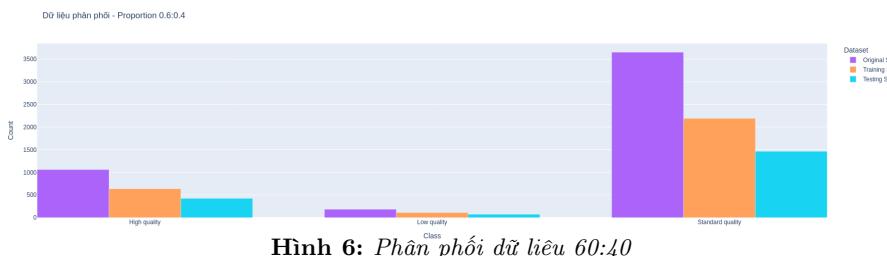


Hình 4: Phân phối dữ liệu 90:10

- Wine Quality Dataset



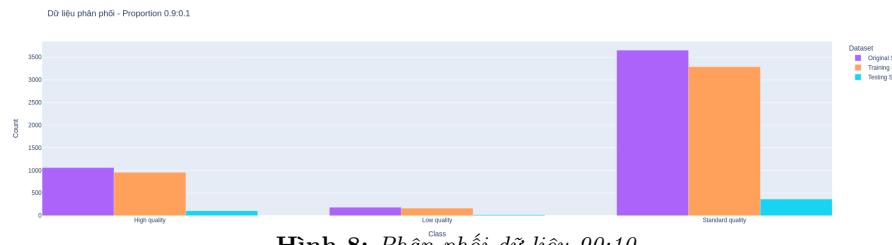
Hình 5: Phân phối dữ liệu 40:60



Hình 6: Phân phối dữ liệu 60:40

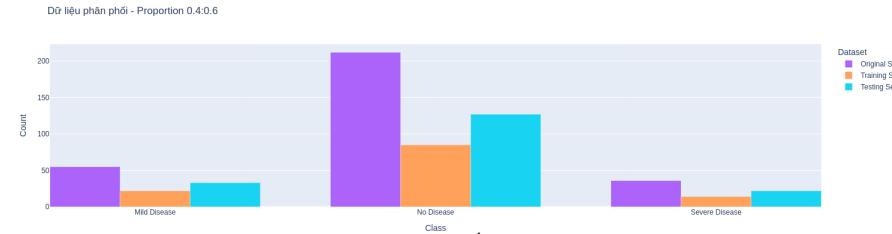


Hình 7: Phân phối dữ liệu 80:20

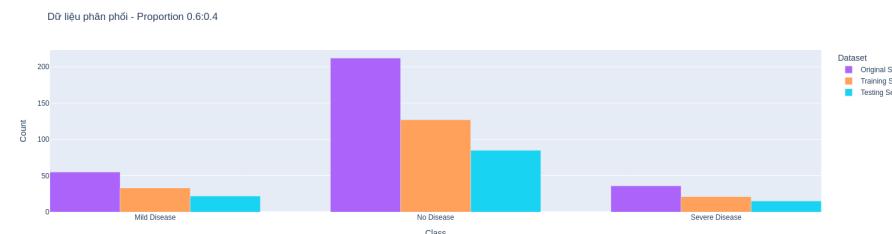


Hình 8: Phân phối dữ liệu 90:10

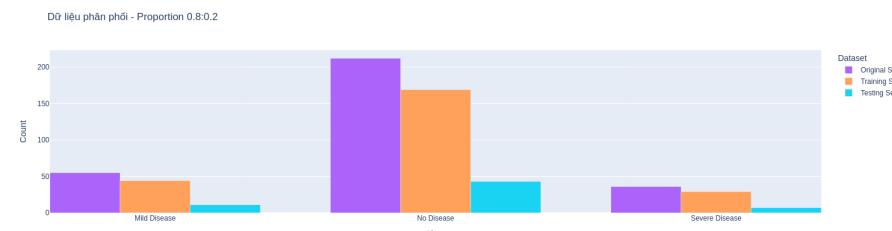
- Heart Disease Dataset



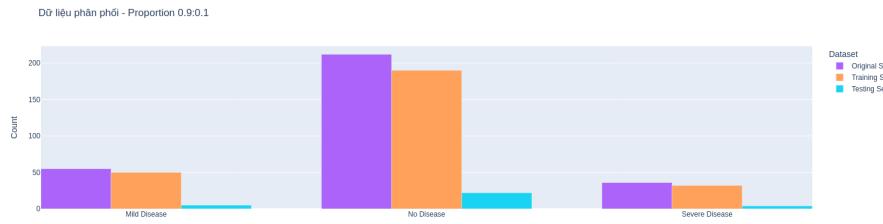
Hình 9: Phân phối dữ liệu 40:60



Hình 10: Phân phối dữ liệu 60:40



Hình 11: Phân phối dữ liệu 80:20



Hình 12: Phân phối dữ liệu 90:10

2 Xây dựng mô hình cây quyết định

2.1 Xây dựng và huấn luyện mô hình với các tỷ lệ khác nhau

Trong phần này, chúng em sẽ xây dựng và huấn luyện các mô hình cây quyết định (decision tree classifiers) cho các tỷ lệ chia tập huấn luyện/kiểm tra khác nhau. Để giải quyết bài toán này, chúng em sử dụng hàm `building_the_decision_tree()`.

Hàm `building_the_decision_tree()` được thiết kế nhằm đơn giản hóa quy trình tạo và huấn luyện các mô hình cây quyết định trên nhiều tập dữ liệu với các tỷ lệ chia huấn luyện/kiểm tra khác nhau. Hàm nhận đầu vào là một danh sách các tập dữ liệu, trong đó mỗi tập đã được chia sẵn thành tập huấn luyện và tập kiểm tra, cùng với tỷ lệ chia tương ứng.

Hàm sẽ lần lượt duyệt qua các tập dữ liệu và tỷ lệ chia tương ứng, huấn luyện một mô hình cây quyết định cho mỗi tập dữ liệu. Tiêu chí được sử dụng để chia nhánh trong cây là *entropy*. Để đảm bảo tính nhất quán và khả năng tái hiện, một *random state* cố định được áp dụng cho từng mô hình. Sau khi huấn luyện, mỗi mô hình được lưu trữ vào ổ đĩa với tên tệp phản ánh tỷ lệ chia tập huấn luyện/kiểm tra cụ thể, giúp dễ dàng nhận diện.

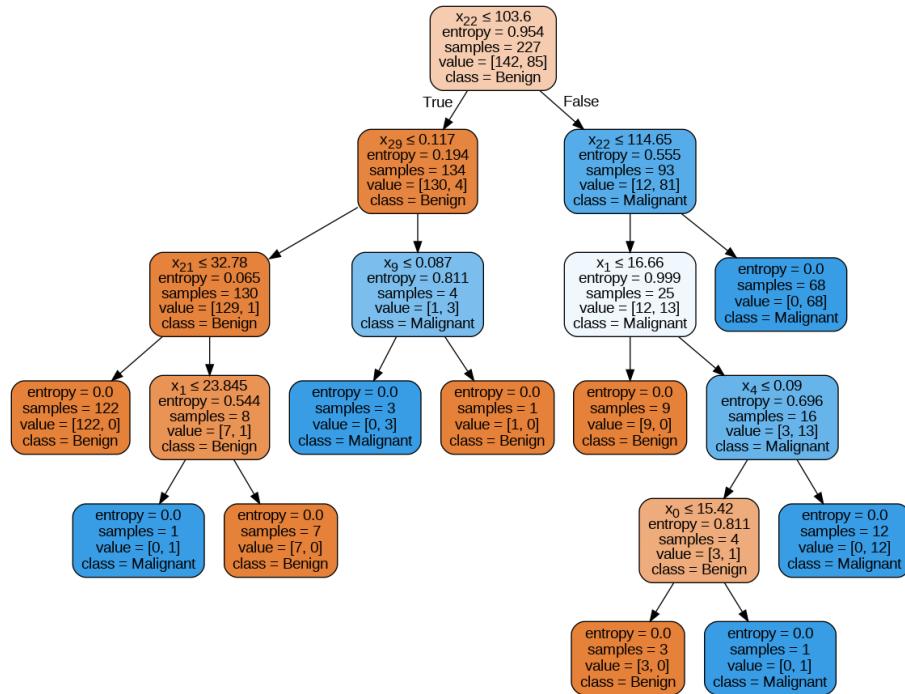
Cuối cùng, hàm trả về danh sách các mô hình đã được huấn luyện, sẵn sàng để phân tích sâu hơn hoặc triển khai vào các ứng dụng thực tế.

2.2 Trực quan hóa cây quyết định

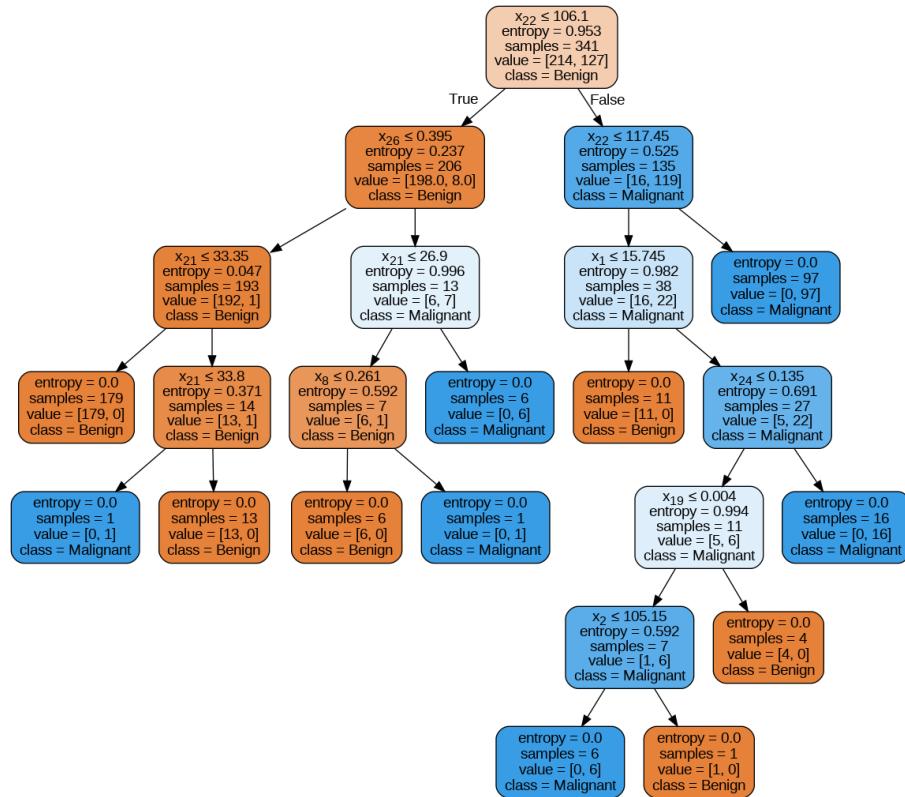
Sau khi huấn luyện, mỗi mô hình cây quyết định được trực quan hóa bằng thư viện `graphviz`, cung cấp một cái nhìn rõ ràng về quá trình ra quyết định mà mô hình đã học được. Việc trực quan hóa thể hiện các yếu tố như tiêu chí chia nhánh tại các nút, nhãn lớp và các đường đi quyết định, giúp cải thiện khả năng giải thích và hỗ trợ phân tích mô hình.

Các biểu đồ trực quan này được lưu dưới dạng tệp PNG cho từng tỷ lệ chia tập huấn luyện và kiểm tra, tạo điều kiện thuận lợi để so sánh và phân tích giữa các kích thước tập dữ liệu khác nhau. Cách tiếp cận có hệ thống này cho phép các nhà nghiên cứu và người thực hành thu được những hiểu biết có giá trị về hành vi ra quyết định của các mô hình đã được huấn luyện, cũng như đánh giá hiệu quả của chúng trong việc xử lý tập dữ liệu đã cho.

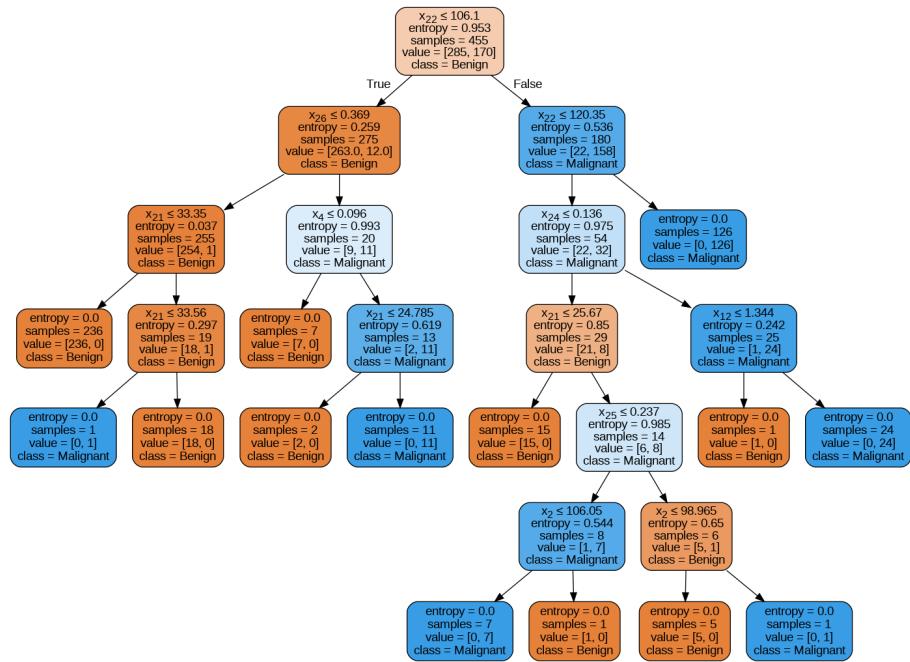
- Breast Cancer Dataset



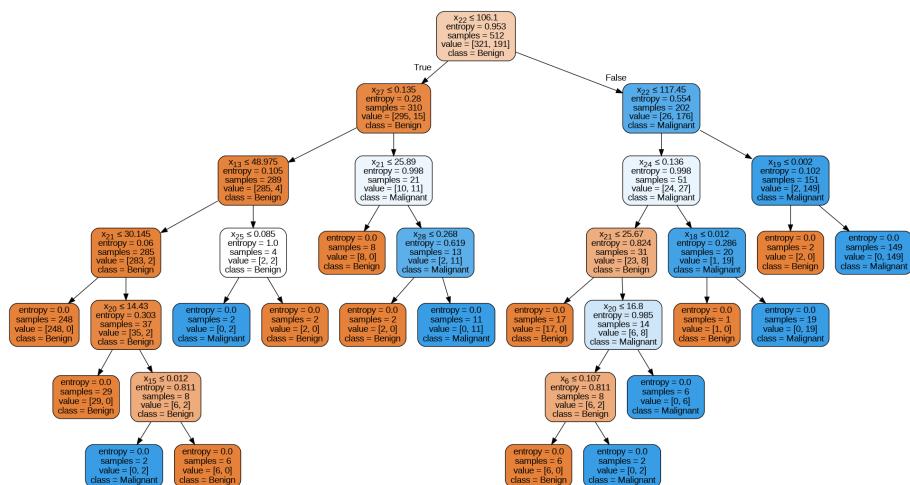
Hình 13: Cây quyết định với tỷ lệ 40_60 lưu vào Dataset 1/split_ratios_graphvizgraph_model_40_60.png



Hình 14: Cây quyết định với tỷ lệ 60_40 lưu vào Dataset 1/split_ratios_graphvizgraph_model_60_40.png



Hình 15: Cây quyết định với tỷ lệ 80_20 lưu vào Dataset 1/split_ratios_graphvizgraph_model_80_20.png



Hình 16: Cây quyết định với tỷ lệ 90_10 lưu vào Dataset 1/split_ratios_graphvizgraph_model_90_10.png

• Wine Quality Dataset

Cây quyết định với tỷ lệ 40_60 lưu vào Dataset 2/split_ratios_graphviz/graph_model_40_60.png
Cây quyết định với tỷ lệ 60_40 lưu vào Dataset 2/split_ratios_graphviz/graph_model_60_40.png
Cây quyết định với tỷ lệ 80_20 lưu vào Dataset 2/split_ratios_graphviz/graph_model_80_20.png
Cây quyết định với tỷ lệ 90_10 lưu vào Dataset 2/split_ratios_graphviz/graph_model_90_10.png

• Heart Disease Dataset

Cây quyết định với tỷ lệ 40_60 lưu vào Dataset 3/split_ratios_graphviz/graph_model_40_60.png
Cây quyết định với tỷ lệ 60_40 lưu vào Dataset 3/split_ratios_graphviz/graph_model_60_40.png
Cây quyết định với tỷ lệ 80_20 lưu vào Dataset 3/split_ratios_graphviz/graph_model_80_20.png
Cây quyết định với tỷ lệ 90_10 lưu vào Dataset 3/split_ratios_graphviz/graph_model_90_10.png



3 Đánh giá mô hình cây quyết định

3.1 Classification report và Confusion matrix

Để diễn giải classification report và confusion matrix từ `sklearn.metrics`, chúng ta sẽ nói về định nghĩa của chúng bởi vì classification report và confusion matrix đóng vai trò quan trọng trong việc đánh giá hiệu quả của một mô hình phân loại.

3.1.1 Classification report

Classification Report cung cấp các chỉ số chính để đánh giá hiệu năng của mô hình, bao gồm:

- **Precision:** Đo lường tỷ lệ các dự đoán đúng thuộc một lớp cụ thể so với tổng các dự đoán thuộc lớp đó. Precision cao nghĩa là mô hình ít nhầm lẫn với các lớp khác.

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

- **Recall:** Đo lường khả năng phát hiện đúng các trường hợp thực sự thuộc về một lớp cụ thể. Recall cao nghĩa là mô hình nhận diện được hầu hết các trường hợp đúng.

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

- **F1-Score:** Là trung bình điều hòa giữa Precision và Recall, giúp cân bằng giữa hai yếu tố này.

$$F_1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

- **Support:** Số lượng mẫu thực tế thuộc từng lớp trong tập dữ liệu.

Confusion Matrix là biểu diễn dưới dạng bảng tóm tắt các dự đoán của mô hình trên tập kiểm tra. Ma trận này cung cấp cái nhìn chi tiết về các dự đoán đúng và sai:

- **True Positives (TP - Dương tính đúng):** Số lượng instance được phân loại *đúng* là lớp dương tính.
- **True Negatives (TN - Âm tính đúng):** Số lượng instance được phân loại *đúng* là lớp âm tính.
- **False Positives (FP - Dương tính sai):** Số lượng instance bị phân loại *sai* là lớp dương tính (Lỗi Loại I).
- **False Negatives (FN - Âm tính sai):** Số lượng instance bị phân loại *sai* là lớp âm tính (Lỗi Loại II).

Ma trận nhầm lẫn thường được biểu diễn dưới dạng bảng 2x2 (cho bài toán phân loại nhị phân):

Predicted / Thực tế	Positive (Dự đoán là Đúng)	Negative (Dự đoán là Sai)
Positive (Thực tế là Đúng)	True Positive (TP)	False Negative (FN)
Negative (Thực tế là Sai)	False Positive (FP)	True Negative (TN)

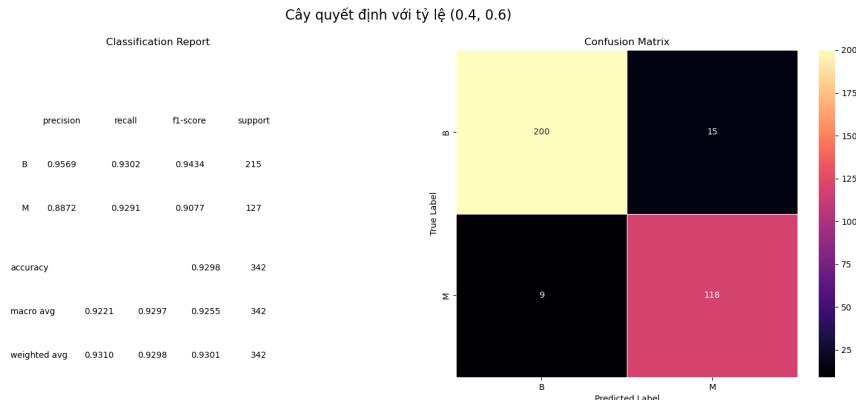
Bảng 7: Confusion Matrix

Việc kết hợp thông tin từ cả classification report (đã được trình bày trước đó) và confusion matrix cung cấp một đánh giá đầy đủ về hiệu suất của bộ phân loại. Classification report cung cấp một tóm tắt tổng quan về các metric hiệu suất chính, trong khi confusion matrix cung cấp một cái nhìn chi tiết về kết quả dự đoán riêng lẻ. Khi được sử dụng cùng nhau, các công cụ này cung cấp một phân tích sâu về điểm mạnh và điểm yếu của bộ phân loại, hỗ trợ việc ra quyết định sáng suốt và nỗ lực cải thiện mô hình.

3.1.2 Confusion matrix

Bằng cách phân tích ma trận nhầm lẫn, chúng ta có thể tính toán các metric hiệu suất trong báo cáo phân loại.

- Breast Cancer Dataset



Hình 17: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 40/60

$$\text{Accuracy} = \frac{\sum TP + TN}{\sum TP + FP + FN + TN} = \frac{200 + 118}{200 + 118 + 15 + 9} \approx 0.9298 \quad (4)$$

Dối với lớp "Lành tính" (B)

– Precision:

$$\text{Precision} = \frac{\sum TP}{\sum TP + FP} = \frac{200}{200 + 9} \approx 0.9569 \quad (5)$$

– Recall:

$$\text{Recall} = \frac{\sum TP}{\sum TP + FN} = \frac{200}{200 + 15} \approx 0.9302 \quad (6)$$

– F₁-Score:

$$F_1\text{-Score} = 2 \times \frac{0.9569 \times 0.9302}{0.9569 + 0.9302} \approx 0.9434 \quad (7)$$

Dối với lớp "Ác tính" (M)

– Precision:

$$\text{Precision} = \frac{\sum TP}{\sum TP + FP} = \frac{118}{118 + 15} \approx 0.8872 \quad (8)$$

– Recall:

$$\text{Recall} = \frac{\sum TP}{\sum TP + FN} = \frac{118}{118 + 9} \approx 0.9291 \quad (9)$$

– F₁-Score:

$$F_1\text{-Score} = 2 \times \frac{0.8872 \times 0.9291}{0.8872 + 0.9291} \approx 0.9077 \quad (10)$$

Trung bình Macro:

– Precision:

$$\text{MacroPrecision} = \frac{\text{Precision}_B + \text{Precision}_M}{2} = \frac{0.9569 + 0.8872}{2} \approx 0.9221 \quad (11)$$

– Recall:

$$\text{MacroRecall} = \frac{\text{Recall}_B + \text{Recall}_M}{2} = \frac{0.9302 + 0.9291}{2} \approx 0.9297 \quad (12)$$

– F₁-Score:

$$\text{MacroF}_1\text{-Score} = \frac{F1_B + F1_M}{2} = \frac{0.9434 + 0.9077}{2} \approx 0.9255 \quad (13)$$

Trung bình có trọng số (Weighted Average):

– Precision:

$$\text{WeightedPrecision} = \frac{(\text{Precision}_B \times \text{Support}_B) + (\text{Precision}_M \times \text{Support}_M)}{\text{TotalSupport}} \quad (14)$$

$$= \frac{(0.9569 \times 215) + (0.8872 \times 127)}{342} \approx 0.9310 \quad (15)$$

– Recall:

$$\text{WeightedPrecision} = \frac{(\text{Recall}_B \times \text{Support}_B) + (\text{Recall}_M \times \text{Support}_M)}{\text{TotalSupport}} \quad (16)$$

$$= \frac{(0.9302 \times 215) + (0.9291 \times 127)}{342} \approx 0.9298 \quad (17)$$

– F₁-Score:

$$\text{WeightedPrecision} = \frac{(F1_B \times \text{Support}_B) + (F1_M \times \text{Support}_M)}{\text{TotalSupport}} \quad (18)$$

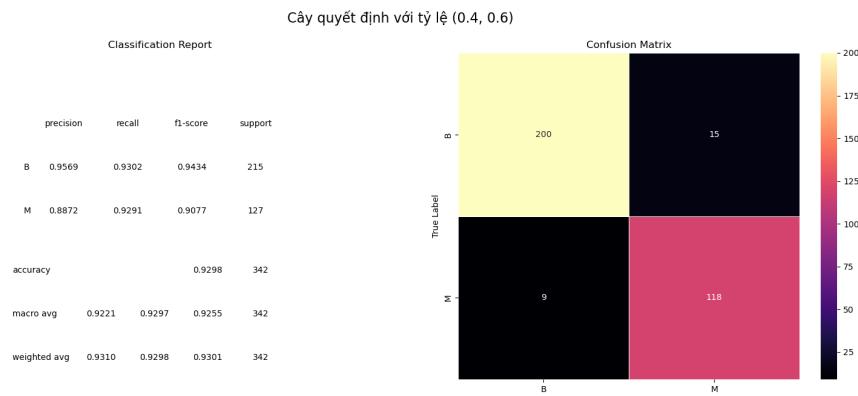
$$= \frac{(0.9434 \times 215) + (0.9077 \times 127)}{342} \approx 0.9301 \quad (19)$$

Cách tính tương tự như vậy cho các số liệu tương ứng của các ma trận khác.

3.2 Nhận xét các biểu đồ trên

3.2.1 Breast Cancer

Tỉ lệ: 40/60



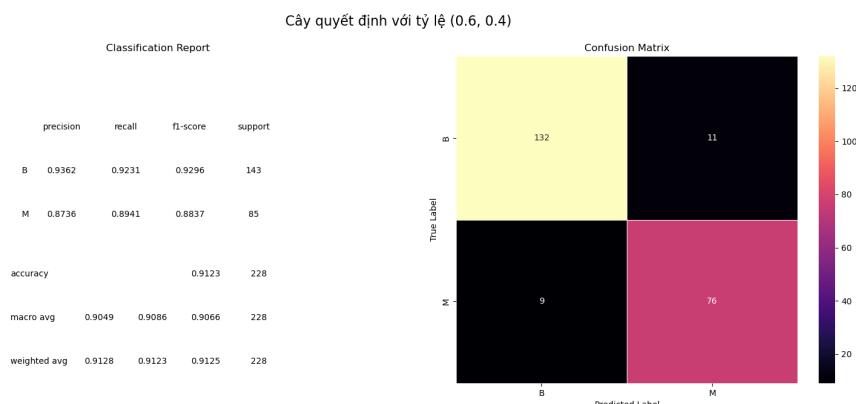
Hình 18: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 40/60

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 92.98%, cho thấy khả năng phân loại hiệu quả các trường hợp thuộc nhiều lớp khác nhau, với tỷ lệ nhãn chính xác cao. Mặc dù hiệu suất này khá ấn tượng, vẫn cần kiểm tra thêm để cải thiện, đặc biệt trong các tình huống mà độ chính xác (precision) hoặc khả năng hồi đáp (recall) có thể chưa đạt mức tối ưu.

- Precision, recall, and F1-score:

- Lớp “Benign” (B): Mô hình đạt độ chính xác cao đối với lớp này, thể hiện rằng số lượng dự đoán dương tính giả (false positives) rất ít. Khoảng 93% các trường hợp lành tính đã được mô hình dự đoán chính xác. Điều này chứng tỏ mô hình có khả năng nhận diện hầu hết các trường hợp lành tính, mặc dù vẫn bỏ sót khoảng 7%. Điểm F1 đạt 94%, cho thấy sự cân bằng tốt giữa độ chính xác và khả năng hồi đáp, minh chứng cho hiệu suất tốt khi phân loại lớp lành tính.
- Lớp “Malignant” (M): Lớp này có độ chính xác và khả năng hồi đáp thấp hơn so với lớp lành tính, nhưng vẫn ở mức cao. Một số nhầm lẫn trong quá trình phân loại đã dẫn đến sự chênh lệch nhẹ giữa các chỉ số. Điểm F1 đạt 90.77%, phản ánh sự cân bằng hợp lý giữa độ chính xác và khả năng hồi đáp, chứng minh khả năng dự đoán ổn định đối với các trường hợp ác tính.

Tỉ lệ: 60/40



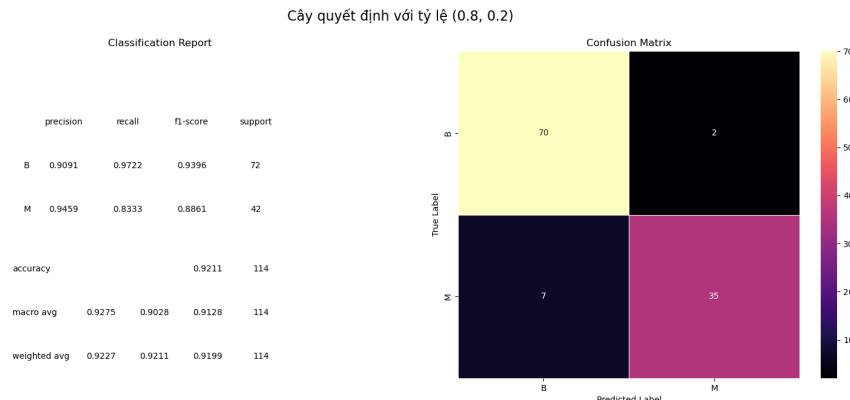
Hình 19: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 60/40

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 91.23%, thể hiện hiệu quả cao trong việc phân loại các trường hợp thuộc mọi lớp. Tuy nhiên, cần xem xét thêm để cải thiện, đặc biệt nếu độ chính xác hoặc khả năng hồi đáp ở một số lớp còn thấp.

- Precision, recall, and F1-score:

- Lớp “Benign” (B): Mô hình có độ chính xác rất cao, với 93.62% các trường hợp dự đoán lành tính là đúng. Khả năng nhận diện đúng các trường hợp lành tính đạt 92.31%. Điểm F1 cho thấy sự cân bằng tốt giữa độ chính xác và khả năng hồi đáp, khẳng định hiệu suất mạnh mẽ khi phân loại lớp này.
- Lớp “Malignant” (M): Độ chính xác đạt 87.36%, thấp hơn một chút so với lớp lành tính, nhưng khả năng hồi đáp cao với 94.12% các trường hợp ác tính được xác định chính xác. Điểm F1 chỉ ra rằng mô hình bỏ sót rất ít trường hợp, với tỷ lệ dương tính và âm tính giả thấp.

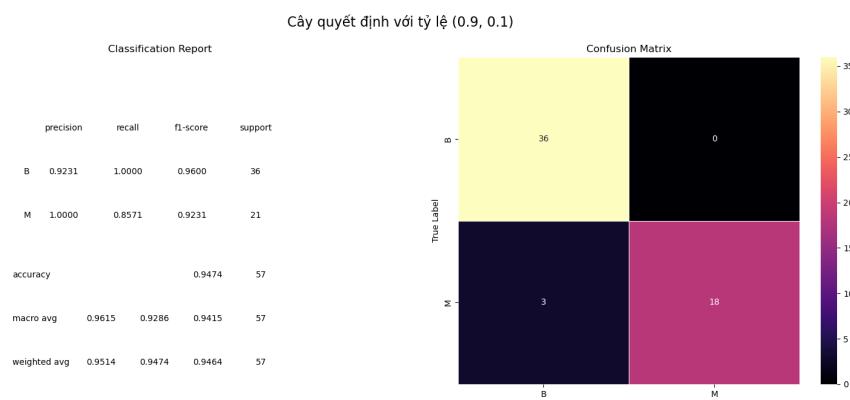
Tỉ lệ: 80/20



Hình 20: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 80/20

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 92.11%, thể hiện khả năng phân loại chính xác cao trên tất cả các lớp.
- Precision, recall, and F1-score:
 - Lớp “Benign” (B): Mô hình dự đoán đúng 90.91% các trường hợp lành tính và xác định chính xác 97.22% các trường hợp lành tính thực tế. Điểm F1 phản ánh sự cân bằng tốt giữa độ chính xác và khả năng hồi đáp, khẳng định hiệu suất mạnh mẽ khi phân loại lớp này.
 - Lớp “Malignant” (M): Độ chính xác đạt 94.59%, cao hơn một chút so với lớp lành tính. Tuy nhiên, khả năng hồi đáp thấp hơn, với 83.33% các trường hợp ác tính được xác định đúng. Điểm F1 cho thấy mô hình vẫn hoạt động tốt, với tỷ lệ dương tính giả và âm tính giả thấp.

Tỉ lệ: 90/10



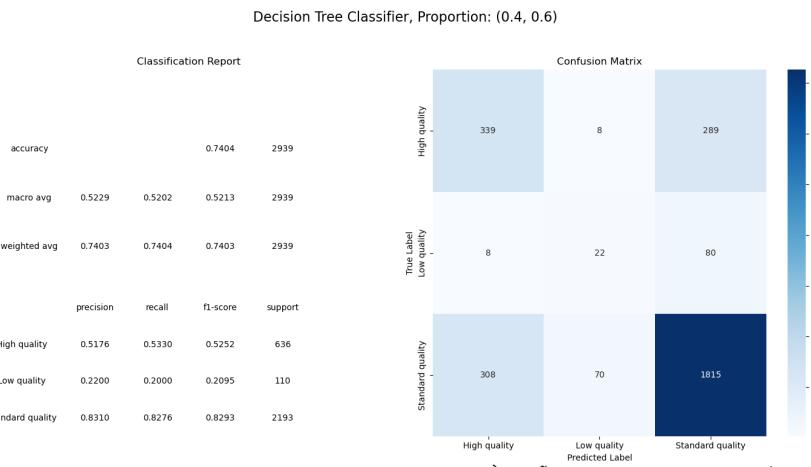
Hình 21: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 90/10

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 94.74%, chứng tỏ khả năng phân loại chính xác cao trên tất cả các lớp.
- Precision, recall, and F1-score:
 - Lớp “Benign” (B): Mô hình đạt điểm độ chính xác và khả năng hồi đáp rất cao. Điểm F1 thể hiện sự cân bằng tốt giữa độ chính xác và khả năng hồi đáp, khẳng định hiệu suất mạnh mẽ khi dự đoán các trường hợp lành tính.
 - Lớp “Malignant” (M): Độ chính xác đạt mức tuyệt đối 100%, với tất cả các dự đoán ác tính đều chính xác. Khả năng hồi đáp cũng cao, với 85.71% các trường hợp ác tính được nhận diện đúng. Điểm F1 đạt 92.31%, thể hiện hiệu quả vượt trội trong việc phân loại các trường hợp ác tính.

Kết luận: Trong bốn tỷ lệ phân chia tập train-test, mô hình phù hợp nhất là mô hình với tỷ lệ 90/10. Mô hình này đạt độ chính xác cao (94.72%) và kích thước lớn của tập huấn luyện giúp mô hình học được nhiều đặc điểm của dữ liệu hơn, từ đó nâng cao giá trị và hiệu quả của mô hình.

3.2.2 Wine Quality

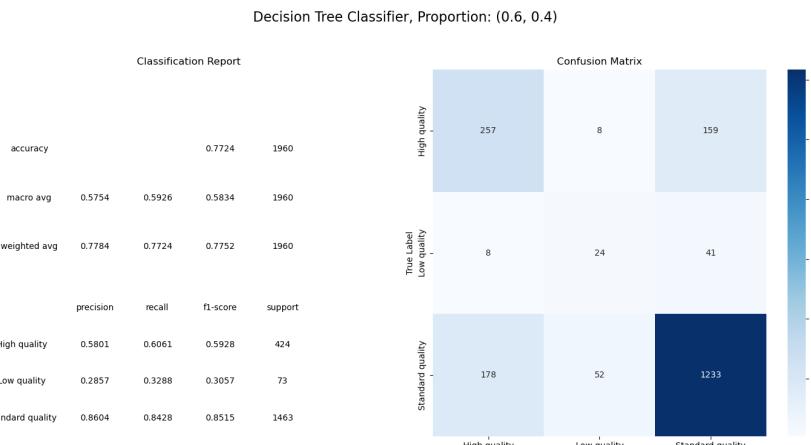
Tỉ lệ: 40/60



Hình 22: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 40/60

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác khoảng 74%, cho thấy khả năng phân loại tương đối tốt đối với các nhóm chất lượng rượu.
- Precision, recall, and F1-score:
 - Chất lượng cao (High Quality): Độ chính xác 52%, độ nhạy 53% và F1-Score 53% cho thấy mô hình chỉ đạt hiệu suất trung bình trong việc phân loại nhóm này. Một số lượng lớn mẫu bị nhầm lẫn sang nhóm Chất lượng tiêu chuẩn (289 mẫu so với 339 mẫu dự đoán đúng).
 - Chất lượng thấp (Low Quality): Độ chính xác 22%, độ nhạy 20% và F1-Score 21% thể hiện hiệu suất phân loại rất kém đối với nhóm này. Mô hình gặp khó khăn trong việc nhận diện các trường hợp thuộc nhóm Chất lượng thấp, với phần lớn mẫu bị nhầm lẫn sang nhóm Chất lượng tiêu chuẩn (80 mẫu so với 22 mẫu dự đoán đúng).
 - Chất lượng tiêu chuẩn (Standard Quality): Đây là nhóm đạt hiệu suất dự đoán tốt nhất với độ chính xác, độ nhạy và F1-Score đều đạt 83%. Mô hình dự đoán chính xác phần lớn các mẫu thuộc nhóm này, đóng góp đáng kể vào độ chính xác tổng thể của mô hình.

Tỉ lệ: 60/40



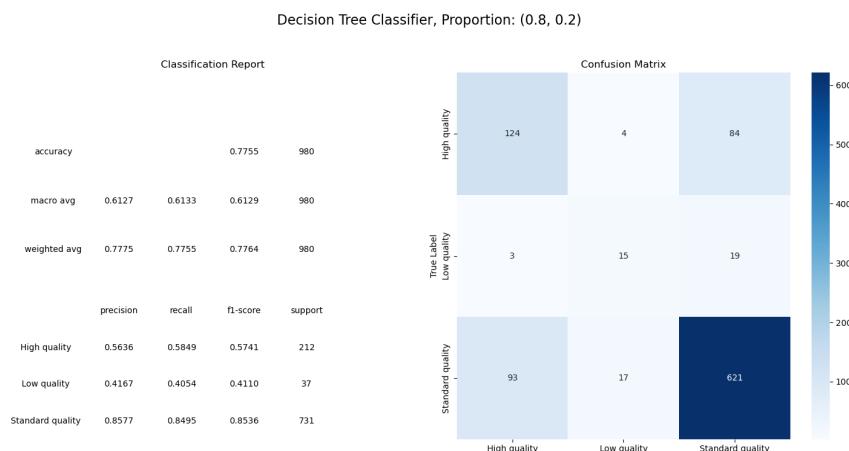
Hình 23: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 60/40

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 77%, cho thấy khả năng phân loại khá tốt, đặc biệt là đối với nhóm Chất lượng tiêu chuẩn (Standard Quality).

- Precision, recall, and F1-score:

- Chất lượng cao (High Quality): Mô hình có độ chính xác 58% và độ nhạy 61%, với F1-Score đạt 59%. Điều này cho thấy mô hình khá hiệu quả trong việc nhận diện nhóm này, mặc dù vẫn có sự nhầm lẫn đáng kể với nhóm Standard Quality, với 159 mẫu được phân loại sai so với 257 mẫu dự đoán đúng.
- Chất lượng thấp (Low Quality): Với độ chính xác chỉ đạt 29%, độ nhạy 33%, và F1-Score 31%, hiệu suất phân loại nhóm này khá yếu. Mô hình gặp khó khăn trong việc phân biệt nhóm Low Quality, dẫn đến việc nhiều mẫu bị nhầm với nhóm Standard Quality. Đặc biệt, số lượng mẫu thuộc nhóm này không nhiều, gây khó khăn cho mô hình trong việc học và phân loại chính xác.
- Chất lượng tiêu chuẩn (Standard Quality): Mô hình đạt độ chính xác 86%, độ nhạy 84%, và F1-Score 85% cho nhóm này, cho thấy khả năng phân loại rất tốt. Phần lớn mẫu của nhóm Standard Quality được dự đoán chính xác, giúp nâng cao độ chính xác chung của mô hình.

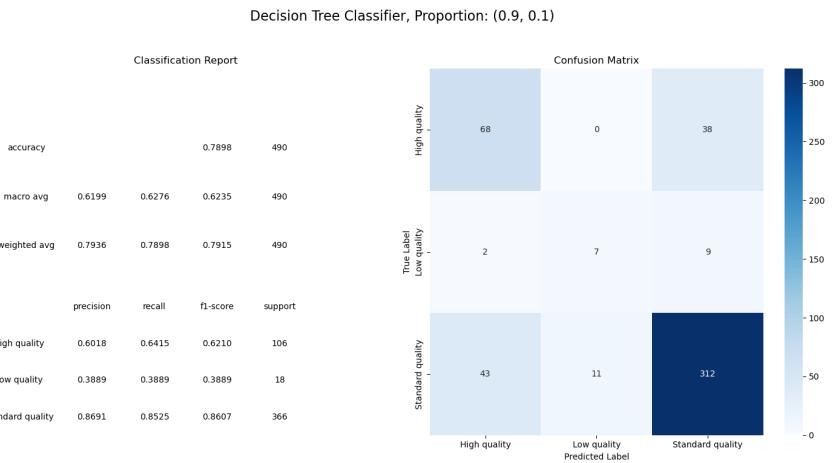
Tỉ lệ: 80/20



Hình 24: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 80/20

- Dộ chính xác (Accuracy): Mô hình đạt độ chính xác 78%, thể hiện hiệu suất khá tốt, đặc biệt trong nhóm Chất lượng Tiêu chuẩn (Standard Quality):
 - Chất lượng cao (High Quality): Độ chính xác 56%, độ nhạy 58%, F1-Score 57%. Dự đoán nhóm này ở mức trung bình, với một số mẫu bị nhầm với nhóm Standard Quality (84 mẫu nhầm so với 124 mẫu đúng).
 - Chất lượng thấp (Low Quality): Độ chính xác 42%, độ nhạy 41%, F1-Score 41%. Mô hình phân loại nhóm này còn hạn chế, nhưng đã cải thiện so với các bộ dữ liệu trước, phản ánh sự khó khăn trong việc phân biệt do số lượng mẫu ít và đặc trưng khó nhận diện.
 - Chất lượng tiêu chuẩn (Standard Quality): Đây là nhóm có hiệu suất cao nhất, với độ chính xác 86%, độ nhạy 85%, F1-Score 85%, nhờ vào việc dự đoán chính xác phần lớn các mẫu, nâng cao độ chính xác chung của mô hình.

Tỉ lệ: 90/10



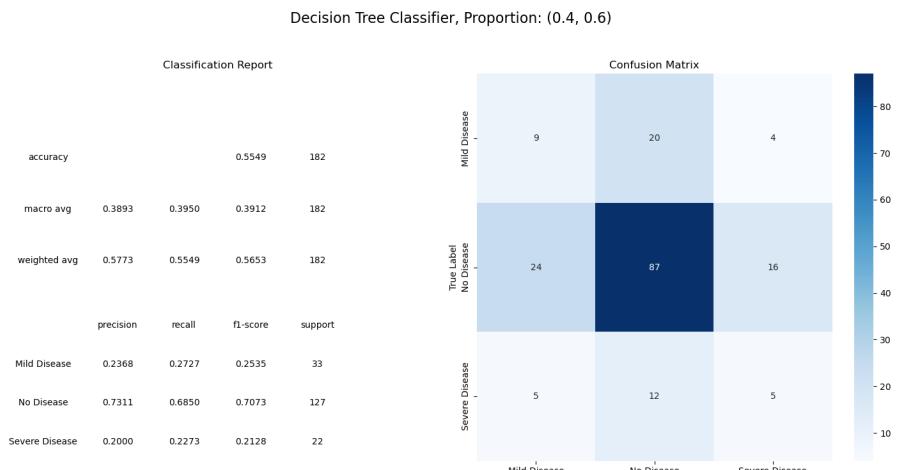
Hình 25: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 90/10

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 79%, thể hiện khả năng phân loại tốt, đặc biệt với nhóm Standard Quality, chiếm phần lớn dữ liệu.
- Precision, recall, and F1-score:
 - Chất lượng cao (High Quality): Độ chính xác 60%, độ nhạy 64%, F1-Score 62%. Mô hình gặp khó khăn trong việc phân loại, với 38 mẫu bị nhầm sang nhóm Standard Quality, so với 68 mẫu dự đoán đúng.
 - Chất lượng thấp (Low Quality): Độ chính xác 39%, độ nhạy 39%, F1-Score 39%. Đây là nhóm khó phân loại nhất, với chỉ 18 mẫu và khả năng dự đoán hạn chế. Hầu hết các nhầm lẫn rơi vào nhóm Standard Quality (9 mẫu, chiếm 50%).
 - Chất lượng tiêu chuẩn (Standard Quality): Nhóm này đạt hiệu suất tốt nhất với độ chính xác 87%, độ nhạy 85%, F1-Score 85%, chỉ có 54 mẫu bị nhầm sang các nhóm khác, đóng góp lớn vào độ chính xác tổng thể của mô hình.

Kết luận Trong bốn tỷ lệ phân chia tập train-test, mô hình phù hợp nhất là mô hình với tỷ lệ 90/10. Mô hình này đạt độ chính xác cao (70%) và kích thước lớn của tập huấn luyện giúp mô hình học được nhiều đặc điểm của dữ liệu hơn, từ đó nâng cao giá trị và hiệu quả của mô hình.

3.2.3 Heart Disease

Tỉ lệ: 40/60



Hình 26: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 40/60

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 55.49%, cho thấy khả năng phân loại chưa cao. Tuy nhiên, cần phân tích thêm để cải thiện hiệu suất, đặc biệt với các lớp có độ chính xác hoặc khả năng hồi đáp thấp.

- Precision, recall, and F1-score:

- Lớp "Mild Disease":

Dộ chính xác đạt 23.68%, rất thấp, cho thấy nhiều trường hợp bệnh nhẹ bị nhầm sang lớp khác. Khả năng hồi đáp chỉ đạt 27.27%, nghĩa là mô hình nhận diện đúng được rất ít trường hợp nhẹ. Điểm F1 là 25.35%, thể hiện hiệu suất phân loại lớp bệnh nhẹ chưa hiệu quả.

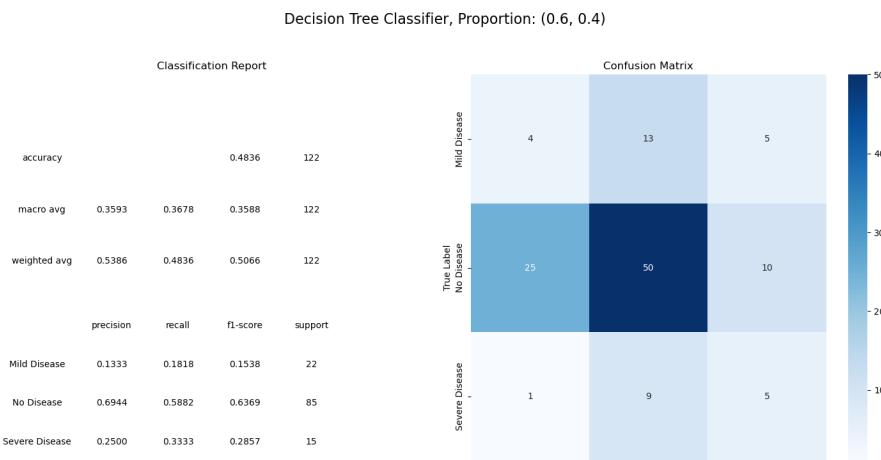
- Lớp "No Disease":

Mô hình đạt độ chính xác 73.11%, cao nhất trong các lớp, với khả năng hồi đáp 68.50%, cho thấy mô hình nhận diện khá tốt các trường hợp không mắc bệnh. Điểm F1 đạt 70.73%, khẳng định sự cân bằng tốt giữa độ chính xác và khả năng hồi đáp cho lớp này.

- Lớp "Severe Disease":

Dộ chính xác chỉ đạt 20.00%, rất thấp, thể hiện rằng nhiều trường hợp bị nhầm lẫn. Khả năng hồi đáp đạt 22.73%, cho thấy mô hình nhận diện không tốt lớp này. Điểm F1 chỉ đạt 21.28%, thể hiện hiệu suất yếu trong việc phân loại các trường hợp thuộc nhóm bệnh nặng.

Tỉ lệ: 60/40



Hình 27: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 60/40

- Dộ chính xác (Accuracy): Mô hình đạt độ chính xác 48.36%, cho thấy khả năng phân loại thấp. Hiệu suất của mô hình cần được cải thiện, đặc biệt là với các lớp "Mild Disease" và "Severe Disease".

- Precision, recall, and F1-score:

- Lớp "Mild Disease":

Dộ chính xác đạt 13.33%, rất thấp, cho thấy nhiều trường hợp thuộc lớp bệnh bị nhầm lẫn. Khả năng hồi đáp chỉ đạt 18.18%, nghĩa là mô hình chỉ nhận diện được một phần nhỏ của các trường hợp thuộc lớp này. Điểm F1 là 15.38%, thể hiện sự yếu kém trong phân loại lớp này.

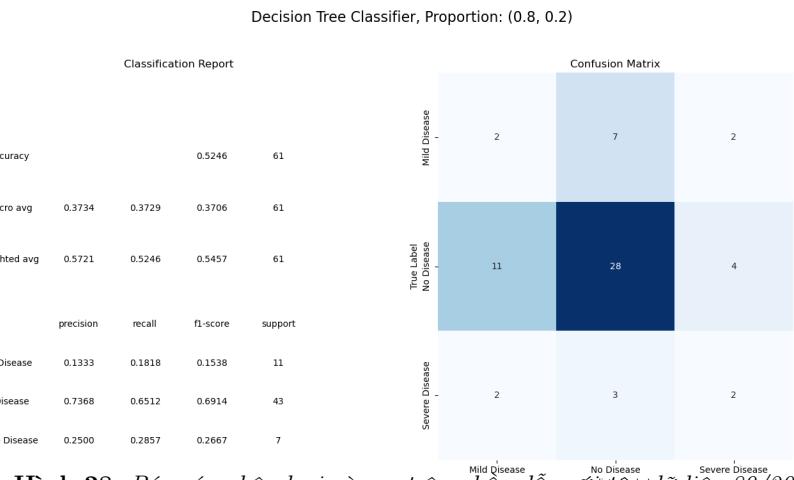
- Lớp "No Disease":

Mô hình đạt độ chính xác 69.44%, tốt nhất trong các lớp, với khả năng hồi đáp 58.82%, cho thấy mô hình hoạt động ổn nhưng vẫn cần cải thiện. Điểm F1 đạt 63.69%, chỉ ra sự cân bằng tương đối giữa độ chính xác và khả năng hồi đáp.

- Lớp "Severe Disease":

Dộ chính xác đạt 25.00%, thấp nhưng tốt hơn lớp "Mild Disease". Khả năng hồi đáp đạt 33.33%, nghĩa là mô hình nhận diện được một số trường hợp thuộc lớp này. Điểm F1 đạt 28.57%, vẫn chưa đủ tốt để phân loại hiệu quả các trường hợp bệnh nặng.

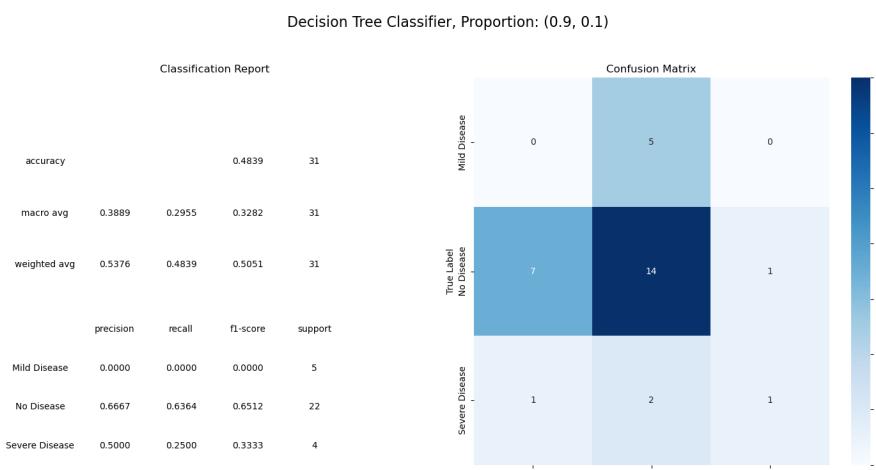
Tỉ lệ: 80/20



Hình 28: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 80/20

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 52.46%, thể hiện sự cải thiện nhẹ so với các thiết lập trước đó. Tuy nhiên, kết quả vẫn chưa đủ tốt để phân loại hiệu quả các lớp, đặc biệt là "Mild Disease" và "Severe Disease".
 - Lớp "Mild Disease": Độ chính xác chỉ đạt 13.33%, cho thấy phần lớn các dự đoán cho lớp này không chính xác. Khả năng hồi đáp (recall) đạt 18.18%, cho thấy mô hình bỏ sót nhiều trường hợp thực sự thuộc lớp này. Điểm F1 là 15.38%, xác nhận hiệu suất thấp khi xử lý lớp này.
 - Lớp "No Disease": Mô hình hoạt động tốt nhất trên lớp này, với độ chính xác đạt 73.68% và khả năng hồi đáp 65.12%, cho thấy phần lớn các trường hợp không có bệnh được nhận diện chính xác. Điểm F1 đạt 69.14%, phản ánh sự cân bằng tương đối giữa độ chính xác và khả năng hồi đáp.
 - Lớp "Severe Disease": Độ chính xác đạt 25.00%, thấp nhưng tốt hơn so với lớp "Mild Disease". Khả năng hồi đáp đạt 28.57%, cho thấy một số trường hợp bệnh nặng được nhận diện đúng. Điểm F1 đạt 26.67%, chỉ ra rằng hiệu suất tổng thể vẫn chưa tốt.

Tỉ lệ: 90/10



Hình 29: Báo cáo phân loại và ma trận nhầm lẫn với tập dữ liệu 90/10

- Độ chính xác (Accuracy): Mô hình đạt độ chính xác 48.39%, cho thấy khả năng phân loại chính xác còn hạn chế trên tất cả các lớp.
- Precision, recall, and F1-score:
 - Lớp "Mild Disease": Mô hình không đạt điểm độ chính xác và khả năng hồi đáp cho lớp này. Điểm F1 cũng bằng 0, thể hiện hiệu suất kém trong việc dự đoán các trường hợp bệnh nhẹ.

- Lớp “No Disease”: Mô hình đạt điểm độ chính xác và khả năng hồi đáp cao nhất cho lớp này. Điểm F1 thể hiện sự cân bằng tốt giữa độ chính xác và khả năng hồi đáp, khẳng định hiệu suất mạnh mẽ khi dự đoán các trường hợp không có bệnh.
- Lớp “Severe Disease”: Độ chính xác và khả năng hồi đáp thấp đối với lớp này. Điểm F1 cũng thấp, thể hiện hiệu quả kém trong việc phân loại các trường hợp bệnh nặng.

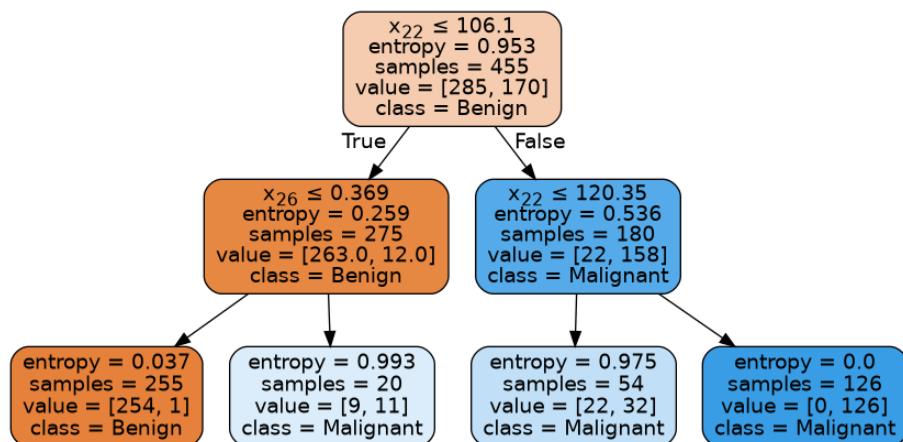
Kết luận Kết luận: Trong bốn tỷ lệ phân chia tập train-test, mô hình phù hợp nhất là mô hình với tỷ lệ 40/60. Mô hình này đạt độ chính xác tương đối (52.46%) và kích thước lớn của tập huấn luyện giúp mô hình học được nhiều đặc điểm của dữ liệu hơn, từ đó nâng cao giá trị và hiệu quả của mô hình.

4 Độ sâu và độ chính xác

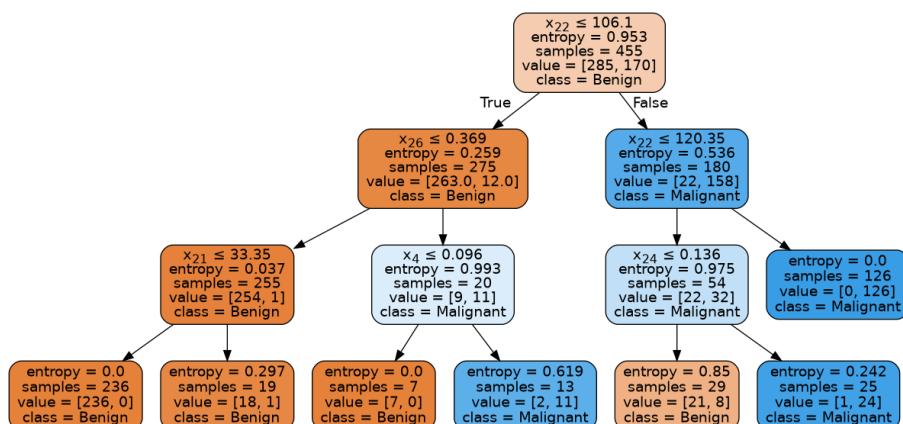
4.1 Cây quyết định với các độ sâu (max_depth) khác nhau

Sau khi đánh giá độ chính xác phân loại ở các độ sâu khác nhau của cây quyết định chia theo tỷ lệ 80/20 với các độ sâu (2, 3, 4, 5, 6, 7 và ‘No limit’), các cấu trúc cây tương ứng được hiển thị bên dưới.

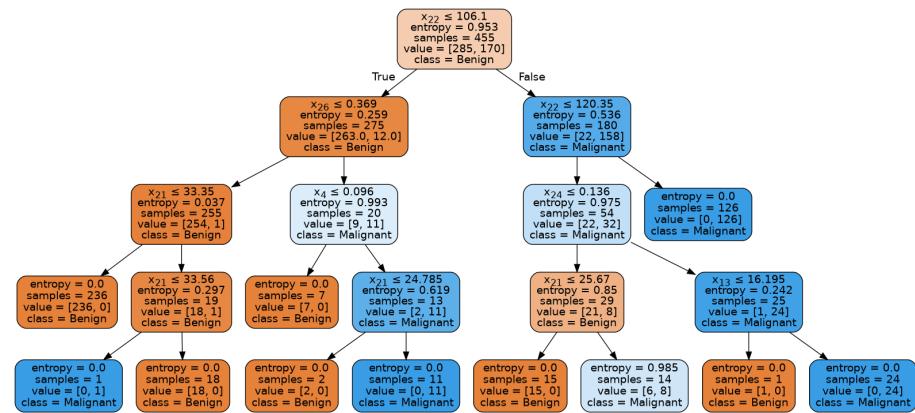
4.1.1 Breast Cancer



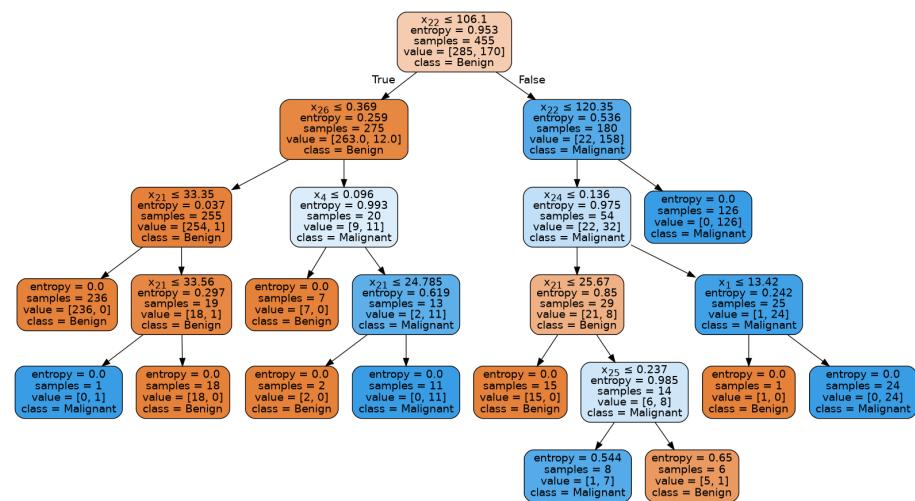
Hình 30: Phân loại cây quyết định với độ sâu 2



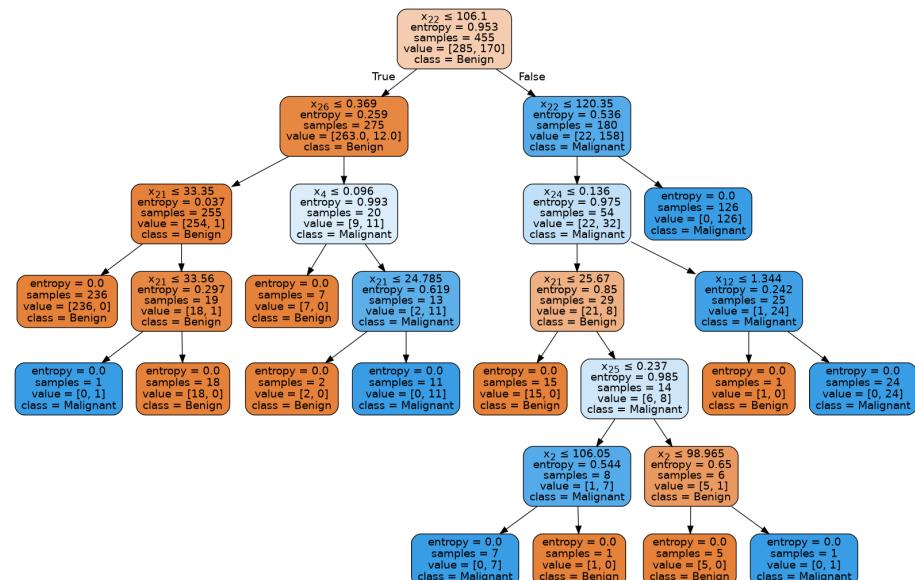
Hình 31: Phân loại cây quyết định với độ sâu 3



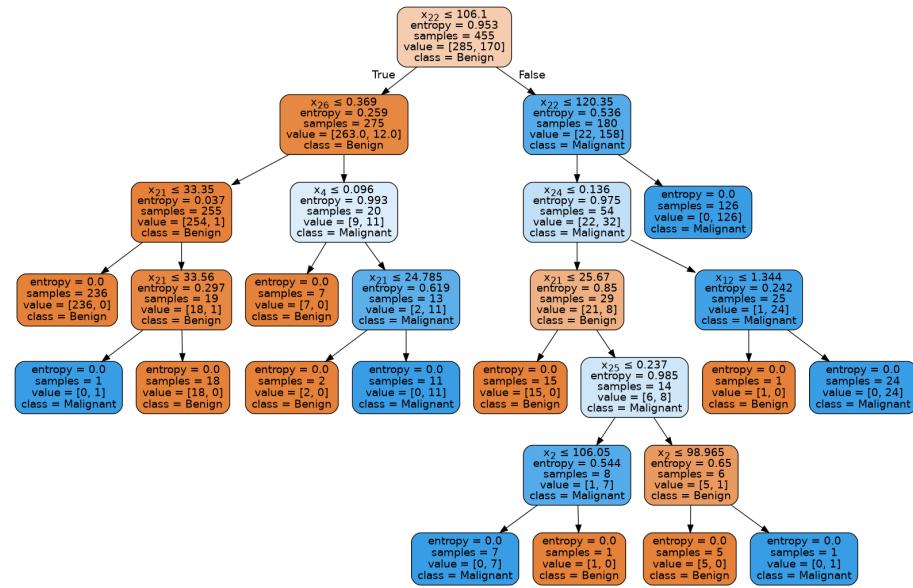
Hình 32: Phân loại cây quyết định với độ sâu 4



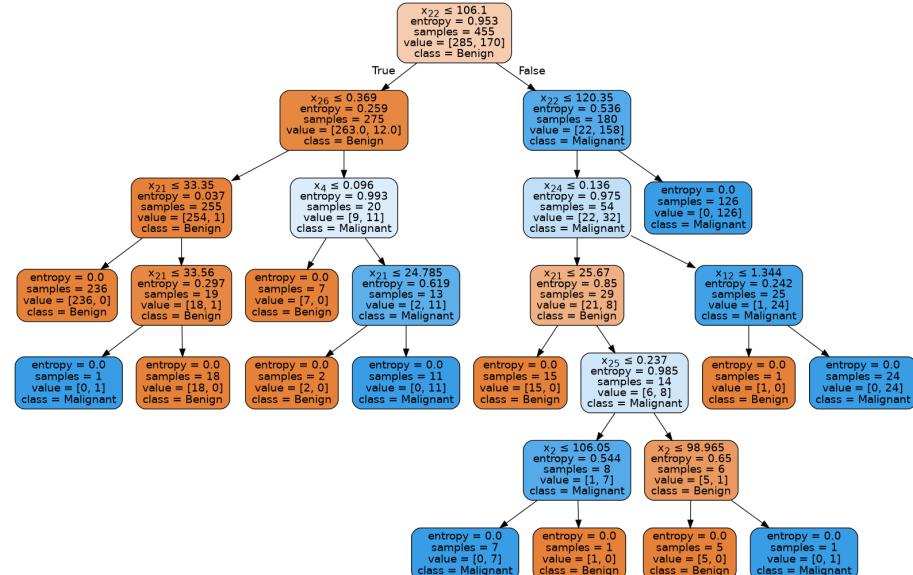
Hình 33: Phân loại cây quyết định với độ sâu 5



Hình 34: Phân loại cây quyết định với độ sâu 6

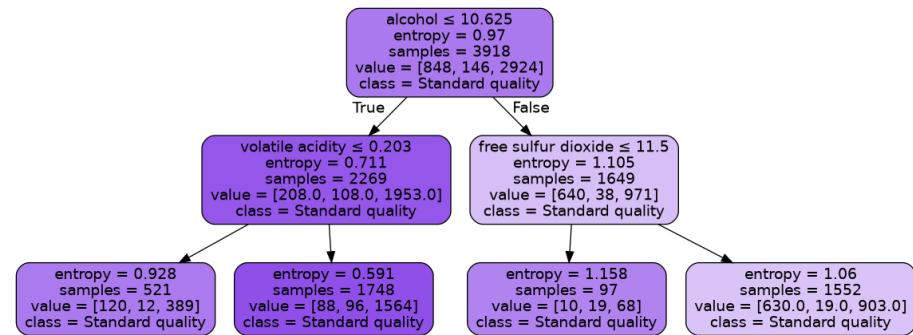


Hình 35: Phân loại cây quyết định với độ sâu 7

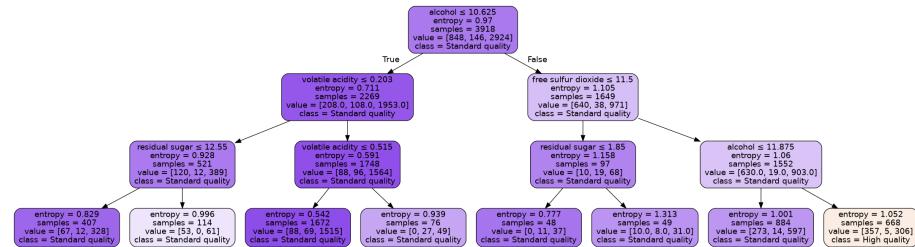


Hình 36: Phân loại cây quyết định với độ sâu không giới hạn

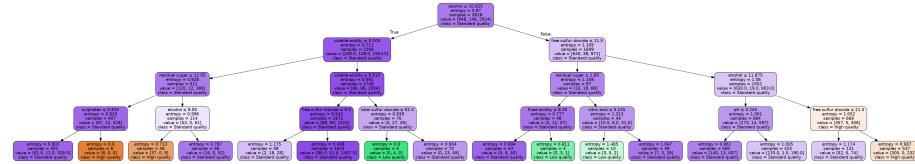
4.1.2 Wine Quality



Hình 37: Phân loại cây quyết định với độ sâu 2



Hình 38: Phân loại cây quyết định với độ sâu 3



Hình 39: Phân loại cây quyết định với độ sâu 4

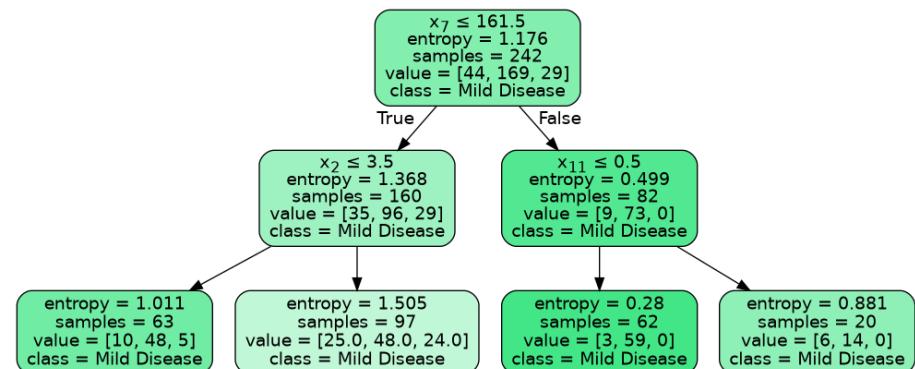
Hình 40: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_5.png

Hình 41: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_6.png

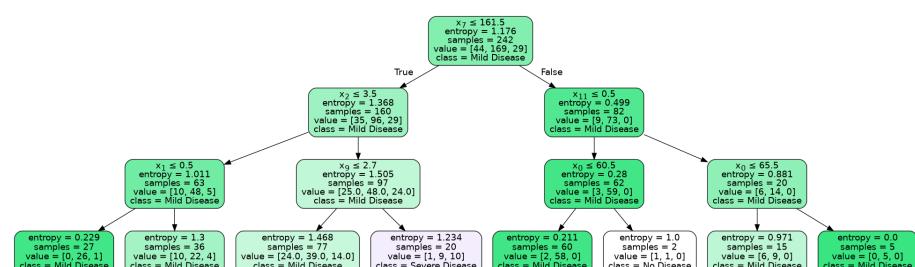
Hình 42: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_7.png

Hình 43: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_8.png

4.1.3 Heart Disease



Hình 44: Phân loại cây quyết định với độ sâu 2



Hình 45: Phân loại cây quyết định với độ sâu 3

Hình 46: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_4png

Hình 47: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_5.png

Hình 48: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_6.png

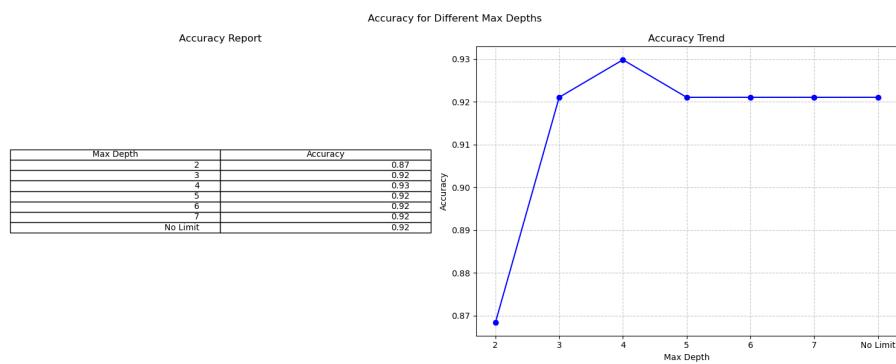
Hình 49: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_7.png

Hình 50: Hình ảnh lưu ở Dataset 2/max_depth_graphviz/graph_model_80_20_depth_8.png

4.2 So sánh và nhận xét ảnh hưởng độ sâu

Các biểu đồ dưới đây minh họa độ chính xác của cây quyết định khi sử dụng tỷ lệ phân chia tập dữ liệu huấn luyện và kiểm tra là 80/20, với các giá trị độ sâu (max_depth) khác nhau.

4.2.1 Breast Cancer

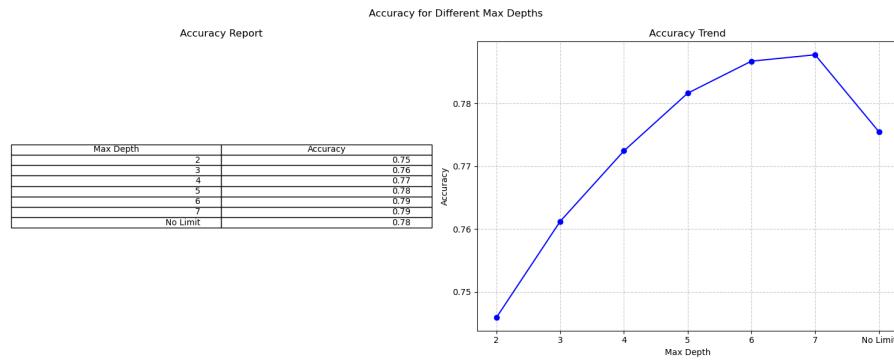


Hình 51: Dánh giá độ chính xác của các chiều sâu

Nhận xét bảng và biểu đồ trên

- Hiệu suất tổng thể: Mô hình đạt độ chính xác cao nhất là 93% khi sử dụng độ sâu tối đa là 4. Đây là mức hiệu suất tối ưu, phản ánh khả năng cân bằng giữa việc tạo ranh giới quyết định phức tạp và tránh tình trạng Overfitting.
- Ảnh hưởng của giá trị max_depth:
 - Độ sâu tối đa 2: Hiệu suất thấp nhất với độ chính xác 86,84%, do mô hình quá đơn giản và không thể nắm bắt đầy đủ các đặc trưng dữ liệu.
 - Độ sâu tối đa 3: Độ chính xác tăng đáng kể lên 92,11%, cho thấy khả năng mô hình cải thiện khi được phép tạo ranh giới phức tạp hơn.
 - Độ sâu tối đa 4: Hiệu suất đạt đỉnh với độ chính xác 92,98%, chứng minh rằng đây là giá trị tối ưu cho mô hình.
 - Độ sâu từ 5 trở lên: Độ chính xác giảm nhẹ và ổn định ở mức 92,11%. Điều này phản ánh rằng việc tăng thêm độ sâu không mang lại lợi ích và có thể dẫn đến hiện tượng quá khớp (overfitting).
- Hiệu suất ổn định: Khi giá trị max_depth vượt quá 4, hiệu suất của mô hình không cải thiện mà duy trì ổn định ở mức 92,11%. Sự ổn định này cho thấy mô hình không thu được thêm thông tin từ việc tăng độ phức tạp, trong khi nguy cơ mất khả năng tổng quát hóa trên dữ liệu mới tăng lên.

4.2.2 Wine Quality

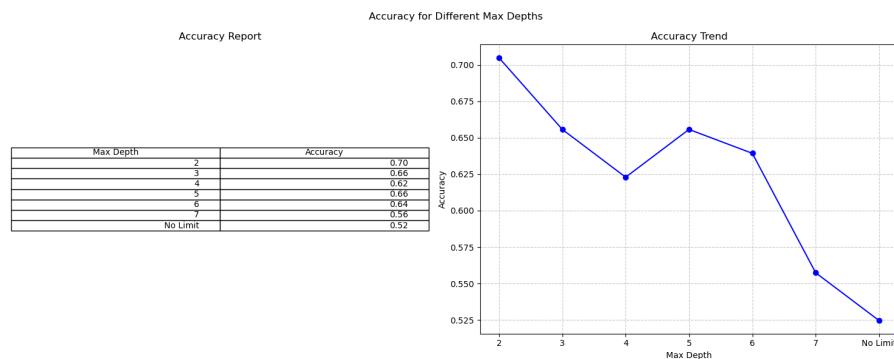


Hình 52: Dánh giá độ chính xác của các chiều sâu

Nhận xét bảng và biểu đồ trên

- Hiệu suất tổng thể: mô hình đạt độ chính xác cao nhất là 79% khi sử dụng giá trị max_depth = 7. Các giá trị max_depth khác dao động từ 75% đến 79%, cho thấy mô hình có hiệu suất khá ổn định khi thay đổi giá trị này.
- Ảnh hưởng của giá trị max_depth:
 - Khi max_depth = None, độ chính xác đạt 78%, cho thấy cây quyết định không giới hạn độ sâu vẫn có thể hoạt động tốt. Tuy nhiên, cây có thể bị overfitting nếu chia nhỏ dữ liệu quá mức, dẫn đến mất khả năng tổng quát hóa.
 - Khi max_depth = 2, độ chính xác thấp nhất đạt 75%, chứng tỏ cây có độ sâu quá nhỏ và không đủ khả năng nhận diện các đặc trưng quan trọng trong dữ liệu phức tạp như Wine Quality.
 - Dối với các giá trị max_depth từ 3 đến 7, độ chính xác tăng dần từ 76% (ở max_depth = 3) lên 79% (ở max_depth = 7), đạt hiệu suất cao nhất tại max_depth = 7.
- Sau khi max_depth đạt 5, độ chính xác duy trì ổn định ở mức rất cao, từ 78% đến 79%. Điều này cho thấy mô hình yêu cầu một cây quyết định có độ sâu trung bình để đạt hiệu quả tốt. Sự gia tăng độ sâu, như khi max_depth = None, không cải thiện đáng kể mà có thể gây Overfitting.

4.2.3 Heart Disease



Hình 53: Dánh giá độ chính xác của các chiều sâu

Nhận xét bảng và biểu đồ trên

- Hiệu suất tổng thể: Mô hình đạt độ chính xác cao nhất là 70% ở độ sâu 2 và hiệu suất giảm dần khi độ sâu tăng. Điều này phản ánh hiện tượng overfitting khi mô hình trở nên quá phức tạp.
- Ảnh hưởng của giá trị max_depth:
 - Độ sâu tối đa 2: mô hình đạt độ chính xác cao nhất (70%), cho thấy rằng cây nông hơn thì có khả năng khai thác tốt hơn cho dữ liệu kiểm tra.



- Độ sâu tối đa 3, 4, 5, 6: độ chính xác của mô hình ở những độ sâu tối đa này dao động từ 62% đến 66%, và mặc dù không giảm mạnh như ở độ sâu tối đa 7 và no limit nhưng điều này cho thấy rằng các độ sâu lớn hơn 2 không thực sự cải thiện hiệu suất.
 - Ở độ sâu tối đa 7, no limit: độ chính xác giảm xuống 2 mức thấp nhất (56% và 52%), điều này cho thấy mô hình trở nên quá phức tạp và bị overfitting.
- Kết luận: Mô hình không ổn định vì độ chính xác thay đổi đáng kể khi điều chỉnh giá trị max_depth. Cụ thể, hiệu suất dao động từ 70% ở độ sâu 2 xuống chỉ còn 52% khi ở độ sâu no limit, cho thấy mô hình dễ bị ảnh hưởng bởi độ phức tạp, thiếu khả năng duy trì độ chính xác ổn định trên các giá trị tham số khác nhau.

5 Phân tích và so sánh các bộ dữ liệu

5.1 Phân tích đặc điểm từng bộ dữ liệu

Breast Cancer Dataset:

- **Loại bài toán:** Phân loại nhị phân.
- **Số lượng mẫu:** Gồm 569 mẫu.
- **Số lượng đặc trưng:** 30, bao gồm các thông tin liên quan đến kích thước, cấu trúc và hình dạng của khối u.
- **Mục tiêu:** Phân loại các khối u thành hai nhóm: lành tính (*Benign - B*) hoặc ác tính (*Malignant - M*).
- **Đặc điểm nổi bật:** Phân bố dữ liệu giữa hai lớp khá cân bằng, tạo điều kiện thuận lợi để mô hình học được đặc trưng phân biệt rõ ràng giữa các nhóm.

Wine Quality Dataset:

- **Loại bài toán:** Phân loại đa lớp.
- **Số lượng mẫu:** Tổng cộng 4898 mẫu.
- **Số lượng đặc trưng:** 11, tập trung vào các chỉ số hóa học như độ axit, độ pH, và nồng độ cồn.
- **Mục tiêu:** Phân loại chất lượng rượu vào ba nhóm:
 - *Low quality (0-4)*: Nhóm rượu chất lượng thấp.
 - *Standard quality (5-6)*: Nhóm rượu chất lượng trung bình.
 - *High quality (7-10)*: Nhóm rượu chất lượng cao.
- **Đặc điểm nổi bật:** Dữ liệu có sự phân bố không đồng đều, với nhóm *Standard Quality* chiếm phần lớn (xấp xỉ 75%), trong khi nhóm *Low Quality* chỉ chiếm một phần nhỏ (khoảng 3.7%).

Heart Disease Dataset:

- **Loại bài toán:** Phân loại đa lớp.
- **Số lượng mẫu:** 303.
- **Số lượng đặc trưng:** 13, bao gồm các đặc trưng liên quan đến thông tin sức khỏe và kết quả kiểm tra như:
 - Tuổi (**age**)
 - Giới tính (**sex**)
 - Kiểu đau ngực (**cp**)
 - Huyết áp lúc nghỉ (**trestbps**)

- Mức cholesterol (chol)
- Đường huyết lúc đói (fbs)
- Kết quả điện tâm đồ (restecg)
- Nhịp tim tối đa (thalach)
- Cơn đau ngực do gắng sức (exang)
- Độ giảm ST (oldpeak)
- Độ dốc của đoạn ST (slope)
- Số lượng mạch bị tổn thương (ca)
- Thalassemia (thal)

- **Mục tiêu:** Phân loại tình trạng bệnh tim thành 3 nhóm:

- **Không bệnh (No Disease):** 164 mẫu (54.1%).
- **Bệnh nhẹ (Mild Disease):** 88 mẫu (29.0%).
- **Bệnh nặng (Severe Disease):** 51 mẫu (16.8%).

- **Đặc điểm nổi bật:** Tập dữ liệu có sự phân bố rõ ràng giữa các lớp, nhưng không đồng đều. Phần lớn mẫu thuộc lớp "Không bệnh", trong khi lớp "Bệnh nặng" có số lượng nhỏ nhất, điều này đòi hỏi chiến lược xử lý dữ liệu phù hợp khi xây dựng mô hình.

5.2 Phân tích các yếu tố ảnh hưởng đến hiệu suất cây quyết định trên 3 bộ dữ liệu

Để đánh giá hiệu quả của mô hình cây quyết định trên các bộ dữ liệu khác nhau, chúng em tiến hành phân tích các yếu tố tác động như số lượng lớp, đặc trưng và kích thước mẫu. Bảng dưới đây thể hiện kết quả so sánh chi tiết:

Hiệu suất tổng quan (Độ chính xác - Accuracy):

Bộ dữ liệu	Tỉ lệ 40/60	Tỉ lệ 60/40	Tỉ lệ 80/20	Tỉ lệ 90/10
Breast Cancer	92.98%	91.23%	92.11%	94.74%
Wine Quality	73.46%	78.67%	78.36%	78.97%
Heart Disease	55.49%	48.36%	52.46%	48.39%

Bảng 8: So sánh độ chính xác (Accuracy) của các bộ dữ liệu với các tỉ lệ chia khác nhau

Nhận định:

- **Tác động của kích thước tập huấn luyện:**

- Dối với bộ dữ liệu **Wine Quality**, hiệu suất mô hình tăng rõ rệt khi tăng kích thước tập huấn luyện, điều này cho thấy mô hình yêu cầu lượng dữ liệu lớn để học tốt hơn các đặc trưng phức tạp. Nguyên nhân có thể do đây là bài toán phân loại đa lớp (3 lớp).
- Bộ dữ liệu **Breast Cancer** và **Heart Disease** cho thấy ít bị ảnh hưởng bởi kích thước tập huấn luyện hơn, có thể do cả hai đều là bài toán phân loại nhị phân.

- **Hiệu suất cao nhất:**

- Bộ dữ liệu **Breast Cancer** đạt độ chính xác cao nhất (94.74%) ở tỉ lệ chia 90/10, khẳng định mô hình hoạt động rất tốt trên tập dữ liệu này.
- Dối với **Wine Quality**, độ chính xác cao nhất (78.97%) đạt được ở tỉ lệ chia 90/10, nhưng không khác biệt nhiều so với tỉ lệ 80/20.
- Bộ dữ liệu **Heart Disease** đạt độ chính xác cao nhất (87.42%) ở tỉ lệ 80/20, nhưng mức chênh lệch giữa các tỉ lệ không đáng kể.



- **Đặc điểm chung:**

- Các tỷ lệ tập huấn luyện lớn hơn, như 80/20 hoặc 90/10, thường mang lại hiệu suất cao hơn. Điều này do tập huấn luyện lớn giúp mô hình học được nhiều đặc trưng hơn, từ đó nâng cao khả năng dự đoán.
- Tỷ lệ 40/60 hoặc 60/40 thường cho hiệu suất thấp nhất, do tập kiểm thử lớn hơn làm tăng nguy cơ xuất hiện lỗi trong quá trình dự đoán.
- Đối với các bộ dữ liệu có sự cân bằng trong phân phối các lớp, như Breast Cancer Dataset, mô hình thường đạt hiệu suất cao hơn so với các bộ dữ liệu có phân phối lớp mất cân bằng, như Wine Quality Dataset hoặc Heart Disease Dataset.

Ảnh hưởng của số lớp:

- **Breast Cancer Dataset:** Bộ dữ liệu gồm 2 lớp: Benign (B) và Malignant (M), với tỷ lệ phân bố lần lượt là 62.7% cho lớp B và 37.3% cho lớp M. Do chỉ có 2 lớp, cấu trúc phân chia của cây quyết định trở nên đơn giản hơn, giúp mô hình dễ dàng xác định ranh giới giữa các đặc trưng. Hơn nữa, sự phân phối tương đối cân bằng giữa hai lớp đảm bảo mô hình nhận diện các mẫu hiệu quả và hạn chế tình trạng thiên lệch. Điều này lý giải tại sao bộ dữ liệu này thường đạt hiệu suất tổng thể cao hơn.
- **Wine Quality Dataset:** Bộ dữ liệu này bao gồm 3 lớp: Chất lượng Cao (High Quality), Chất lượng Tiêu chuẩn (Standard Quality), và Chất lượng Thấp (Low Quality). Tỷ lệ phân bố các lớp lần lượt là 21.6%, 74.6%, và 3.7%. Do lớp Standard Quality chiếm phần lớn dữ liệu, mô hình thường học tốt các đặc trưng của lớp này, dẫn đến khả năng dự đoán chính xác hơn. Tuy nhiên, lớp Low Quality có tỷ lệ nhỏ (3.7%), làm giảm khả năng dự đoán chính xác đối với lớp này, thậm chí có thể bị bỏ qua, khiến mô hình khó đạt hiệu suất tốt trên toàn bộ tập dữ liệu.
- **Heart Disease Dataset:** Bộ dữ liệu này chứa thông tin về các yếu tố nguy cơ và chẩn đoán bệnh tim mạch từ Cleveland Clinic Foundation. Thuộc tính mục tiêu được phân loại thành ba nhóm: **Không bệnh (No Disease)**, **Bệnh nhẹ (Mild Disease)**, và **Bệnh nặng (Severe Disease)**. Theo báo cáo thực nghiệm:
 - Lớp **Không bệnh (No Disease)**: chiếm tỷ lệ cao nhất, với **164 mẫu (54.1%)**.
 - Lớp **Bệnh nhẹ (Mild Disease)**: chiếm **88 mẫu (29.0%)**.
 - Lớp **Bệnh nặng (Severe Disease)**: chiếm **51 mẫu (16.8%)**.- **Số lượng lớp (3 lớp):** Việc phân loại thành ba nhóm giúp cung cấp thông tin chi tiết hơn về mức độ nghiêm trọng của bệnh tim, hỗ trợ tốt cho việc ra quyết định điều trị.
- **Mức độ cân bằng:** Bộ dữ liệu có sự mất cân bằng đáng kể giữa các lớp, đặc biệt là lớp **Severe Disease** có tỷ lệ thấp nhất (16.8%). Điều này có thể gây khó khăn trong việc nhận diện chính xác các mẫu thuộc lớp thiểu số.
- **Tác động:**
 - * **Ưu điểm:** Lớp **Không bệnh (No Disease)** với tỷ lệ cao (54.1%) giúp mô hình học được đặc trưng phổ biến của nhóm này, dẫn đến hiệu suất cao trong việc dự đoán các trường hợp không mắc bệnh.
 - * **Nhược điểm:** Sự chênh lệch tỷ lệ giữa các lớp có thể khiến mô hình thiên lệch, dẫn đến việc phân loại kém hiệu quả các lớp thiểu số (**Mild Disease** và **Severe Disease**). Do đó, cần áp dụng các phương pháp xử lý dữ liệu mất cân bằng (như oversampling hoặc undersampling) và sử dụng các chỉ số như Precision, Recall, F1-score để đánh giá hiệu suất toàn diện hơn.

Ảnh hưởng của số lượng đặc trưng:

- **Breast Cancer Dataset:**

Bộ dữ liệu bao gồm 30 đặc trưng liên tục, giúp mô hình có khả năng nhận diện các mẫu phức tạp nhờ vào thông tin chi tiết mà các đặc trưng cung cấp. Số lượng đặc trưng lớn hỗ trợ cây quyết định hoạt động hiệu quả trong việc phân tách dữ liệu. Tuy nhiên, việc sử dụng nhiều đặc trưng cũng có thể làm tăng độ phức tạp của cây, dẫn đến nguy cơ *overfitting* trên tập huấn luyện. Điều này đòi hỏi phải kiểm soát tốt tham số `max_depth` để tránh hiện tượng này.



- **Wine Quality Dataset:**

Bộ dữ liệu gồm 11 đặc trưng liên tục liên quan đến các thông số hóa học của rượu, như độ axit, nồng độ cồn, và độ pH. Với số lượng đặc trưng vừa phải, mô hình có thể tập trung vào những yếu tố quan trọng, tránh việc bị quá tải thông tin. Tuy nhiên, nếu có đặc trưng nhiều hoặc không liên quan, hiệu suất của mô hình có thể bị ảnh hưởng tiêu cực.

- **Heart Disease Dataset:**

Bộ dữ liệu gồm 13 đặc trưng với sự kết hợp giữa các đặc trưng liên tục (như tuổi, cholesterol, nhịp tim tối đa) và phân loại (như giới tính, loại đau ngực, thalassemia).

* **Ưu điểm:**

- Sự kết hợp giữa các đặc trưng liên tục và phân loại giúp mô hình khai thác được nhiều nguồn thông tin từ dữ liệu.
- Các đặc trưng liên quan đến y học, như loại đau ngực (*Chest Pain Type*) và nhịp tim tối đa (*Maximum Heart Rate*), cung cấp thông tin quan trọng để cải thiện hiệu suất dự đoán.

* **Thách thức:**

- Việc mã hóa các đặc trưng phân loại đòi hỏi kỹ thuật xử lý phù hợp để tránh mất thông tin.
- Một số đặc trưng, như đường huyết lúc đói (*Fasting Blood Sugar*), có thể ít biến thiên, làm giảm giá trị đóng góp vào mô hình.
- Cần kiểm soát tốt các tham số cây quyết định để tránh hiện tượng *overfitting* do các đặc trưng liên tục.

Ảnh hưởng của số lượng mẫu

- **Breast Cancer Dataset:** Sở hữu 569 mẫu dữ liệu, mô hình cây quyết định thể hiện khả năng phân loại tốt trên các tập dữ liệu được phân chia theo nhiều tỷ lệ khác nhau. Tuy nhiên, kích thước mẫu tương đối nhỏ này cũng khiến mô hình nhạy cảm với các biến động trong tập kiểm tra, đặc biệt khi tập kiểm tra có kích thước nhỏ (ví dụ: tỷ lệ chia 90/10). Trong trường hợp này, dù đạt được độ chính xác cao trên tập kiểm tra, mô hình có thể không phản ánh đúng khả năng tổng quát hóa trên dữ liệu mới.

- **Wine Quality Dataset:** Tập dữ liệu này chứa 4898 mẫu (bao gồm cả Red Wine và White Wine), được xem là kích thước trung bình đến lớn. Khi tỷ lệ chia tập huấn luyện/kiểm tra thấp (ví dụ: 60/40 hoặc 40/60), số lượng mẫu lớn trong tập huấn luyện hỗ trợ mô hình học được các mối quan hệ phức tạp giữa các đặc trưng đầu vào. Tuy nhiên, do sự mất cân bằng lớp nghiêm trọng, các lớp thiểu số (ví dụ: chất lượng cao hoặc chất lượng thấp) có số lượng mẫu ít hơn đáng kể. Sau khi chia dữ liệu, điều này khiến việc nhận diện các lớp thiểu số trở nên khó khăn, làm giảm hiệu suất tổng thể của mô hình trên các lớp này.

- **Heart Disease Dataset:** Bộ dữ liệu này chứa 303 mẫu, được xem là kích thước nhỏ đến trung bình.

* **Ưu điểm:** Số lượng mẫu tương đối nhỏ này, kết hợp với sự phân bố lớp hợp lý giữa các nhóm Không bệnh, Bệnh nhẹ, và Bệnh nặng, cho phép mô hình học được các đặc trưng cơ bản và đạt hiệu suất chấp nhận được ở mức tổng thể trên các tỷ lệ phân chia dữ liệu khác nhau.

* **Nhược điểm:** Kích thước mẫu nhỏ làm tăng nguy cơ mô hình bị ảnh hưởng bởi nhiều và các điểm ngoại lai (outliers) trong dữ liệu. Đặc biệt, khi chia dữ liệu theo tỷ lệ 90/10, tập kiểm tra chỉ còn khoảng 30 mẫu. Điều này không đủ lớn để đánh giá chính xác khả năng tổng quát hóa của mô hình, dẫn tới rủi ro rằng mô hình sẽ không hoạt động tốt khi áp dụng cho dữ liệu thực tế.

6 Tổng kết

Trong dự án này, nhóm nghiên cứu đã xây dựng thành công mô hình cây quyết định cho ba bộ dữ liệu: Breast Cancer, Wine Quality, và Heart Disease. Các kết quả quan trọng bao gồm:



• Chuẩn bị dữ liệu:

- Phân chia dữ liệu theo các tỷ lệ 40/60, 60/40, 80/20, và 90/10, đảm bảo tính đại diện của các lớp bằng phương pháp phân chia ngẫu nhiên phân tầng (stratified splitting).
- Sử dụng các biểu đồ để trực quan hóa dữ liệu, qua đó phát hiện các vấn đề tiềm ẩn như sự mất cân bằng lớp trong bộ dữ liệu Wine Quality và Heart Disease, hỗ trợ cho việc phân tích chuyên sâu.

• Xây dựng và huấn luyện mô hình:

- Triển khai mô hình cây quyết định cho từng bộ dữ liệu, sử dụng độ lợi thông tin (entropy) làm tiêu chí chính để phân chia nút.
- Trực quan hóa cấu trúc cây quyết định, giúp làm rõ quá trình phân loại và dự đoán của mô hình.

• **Dánh giá hiệu suất:** Hiệu suất của mô hình trên từng tập dữ liệu với các tỷ lệ phân chia khác nhau được đánh giá và phân tích chi tiết trong phần 4.2 và 6.2, phụ thuộc vào đặc thù riêng của từng bộ dữ liệu.

• **Ảnh hưởng của độ sâu cây:** Mỗi bộ dữ liệu cho thấy một độ sâu cây tối ưu khác nhau để đạt được hiệu suất cao nhất. Mỗi quan hệ giữa độ sâu cây và hiệu suất mô hình được phân tích cụ thể trong phần 5.2, phụ thuộc vào các yếu tố đặc trưng của từng bộ dữ liệu.