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


fit@hcmus
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Tấn Duy Anh - Bùi Hồng Phúc
Nguyễn Lê Anh Phúc - Hồ Minh Quang

BÁO CÁO ĐỒ ÁN

# CƠ SỞ TRÍ TUỆ NHÂN TẠO

## ĐỒ ÁN 2: CÂY QUYẾT ĐỊNH

**Lớp:** 22_21

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 12 NĂM 2024

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


# BÁO CÁO ĐỒ ÁN
# CƠ SỞ TRÍ TUỆ NHÂN TẠO
## ĐỒ ÁN 2: CÂY QUYẾT ĐỊNH

**Môn học:** Cơ sở trí tuệ nhân tạo
**Mã môn học:** CSC14003
**Lớp:** 22_21
**Giảng viên hướng dẫn:** GS. TS. Lê Hoài Bắc; CN. Nguyễn Thanh Tình
**Các sinh viên tham gia:**



| Họ và tên | Mã số sinh viên |
|---|---|
| Nguyễn Tấn Duy Anh | 22120015 |
| Bùi Hồng Phúc | 22120270 |
| Nguyễn Lê Anh Phúc | 22120276 |
| Hồ Minh Quang | 22120295 |



THÀNH PHỐ HỒ CHÍ MINH, THÁNG 12 NĂM 2024

# Lời cảm ơn

Chúng em xin chân thành càm ơn thầy Lê Hoài Bắc đã hướng dẫn phần lý thuyết trên lớp và thầy Nguyễn Thanh Tình đã hướng dẫn phần đồ án thực hành để nhóm có thể hoàn thành đồ án này. Chúng em rất mong nhận được sự góp ý nếu có sai sót cũng như hạn chế trong quá trình làm đồ án này.

Thành phố Hồ Chí Minh, tháng 12 năm 2024

# Mục lục



# Danh sách hình

# 1 Giới thiệu

## 1.1 Bảng phân công và đánh giá tiến độ công việc

Dưới đây là bảng phân công và đánh giá tiến độ công việc của nhóm:

| Họ và tên | Các công việc phụ trách | Tiến độ công việc |
|---|---|---|
| Nguyễn Tấn Duy Anh | Cài đặt thuật toán BFS | 95% |
| | Viết giới thiệu trò chơi và thuật toán BFS | 95% |
| Bùi Hồng Phúc | Cài đặt thuật toán DFS | 95% |
| | Thiết kế giao diện chương trình | 95% |
| | Viết thuật toán DFS | 100% |
| Nguyễn Lê Anh Phúc | Cài đặt thuật toán A* | 95% |
| | Kiểm thử chương trình | 95% |
| | Viết thuật toán A* | 95% |
| Hồ Minh Quang | Cài đặt thuật toán UCS | 95% |
| | Viết toàn bộ khung báo cáo LaTeX | 100% |
| | Viết thuật toán UCS | 95% |

## 1.2 Sơ lược về cây quyết định

# 2 Tập dữ liệu thứ nhất

## 2.1 Giới thiệu về tập dữ liệu

**Breast Cancer Wisconsin (Diagnostic)**

Đây là một tập dữ liệu nổi tiếng trong lĩnh vực học máy và phân tích dữ liệu y tế, được thiết kế để phân loại khối u vú là **lành tính (Benign)** hoặc **ác tính (Malignant)** dựa trên các đặc trưng số học được trích xuất từ hình ảnh tế bào.

**Thông tin tổng quan về tập dữ liệu**

- **Nguồn gốc**: UCI Machine Learning Repository.
- **Mục đích**: Phân loại khối u vú là *lành tính* (benign) hoặc *ác tính* (malignant).
- **Số mẫu**: 569.
- **Số đặc trưng**: 30 đặc trưng đầu vào (dạng số học):
  1. **Bán kính (radius)**.
  2. **Kết cấu (texture)**.
  3. **Chu vi (perimeter)**.
  4. **Diện tích (area)**.
  5. **Độ nhẵn (smoothness)**.

6. **Độ gọn (compactness)**.
7. **Độ lồi (concavity)**.
8. **Số điểm lõm (concave points)**.
9. **Đối xứng (symmetry)**.
10. **Kích thước phân dạng (fractal dimension)**.

- **Nhãn đầu ra**:
  - **0**: *Lành tính (Benign)*.
  - **1**: *Ác tính (Malignant)*.

**Phân phối lớp**

Tập dữ liệu **Breast Cancer Wisconsin (Diagnostic)** có 569 mẫu, trong đó:

- **Lành tính(Benign)**: 357 mẫu, chiếm 37.3% trong 569 mẫu.
- **Ác tính(Malignant)**: 212 mẫu. chiếm 62.7% trong 569 mẫu.

## 2.2 Tiền xử lý dữ liệu

**Chia dữ liệu huấn luyện và dữ liệu kiểm tra**

- 40% tập huấn luyện, 60% tập kiểm tra.
- 60% tập huấn luyện, 40% tập kiểm tra.
- 80% tập huấn luyện, 20% tập kiểm tra.
- 90% tập huấn luyện, 10% tập kiểm tra.

Quá trình chia dữ liệu được thực hiện bằng cách sử dụng hàm **train_test_split** từ thư viện **sklearn.model_selection**. Hàm này đảm bảo rằng việc chia dữ liệu là ngẫu nhiên và đồng nhất. Ví dụ cách sử dụng:

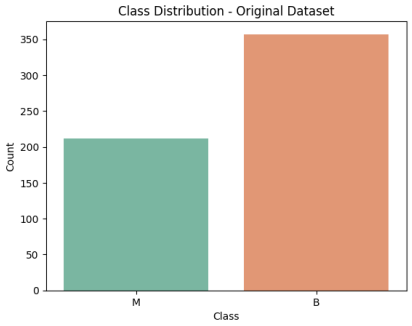
```
from sklearn.model_selection import train_test_split

# Chia dữ liệu với tỷ lệ 80/20
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                    random_state=42)
```

Trong đó:

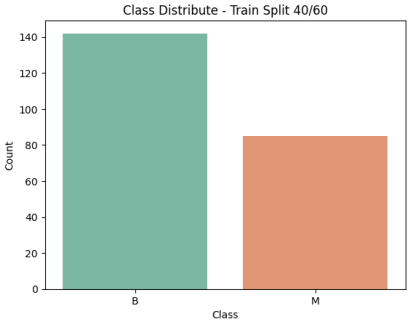
- `X`: Ma trận đặc trưng (features).
- `y`: Nhãn đầu ra (labels).
- `test_size`: Tỷ lệ của tập kiểm tra so với dữ liệu tổng.
- `random_state`: Giá trị cố định để đảm bảo kết quả chia dữ liệu có thể tái lập.

### Tỷ lệ 40/60

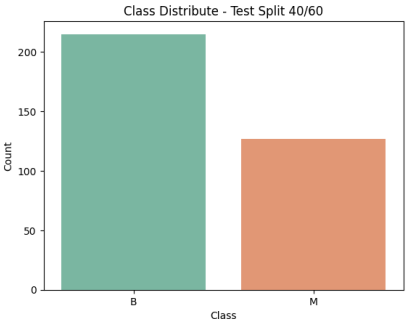
- **Tập huấn luyện**: 227 mẫu (142 Benign, 85 Malignant).
- **Tập kiểm tra**: 342 mẫu (215 Benign, 127 Malignant).




Hình 2.1: Original Dataset




Hình 2.2: Training 40%

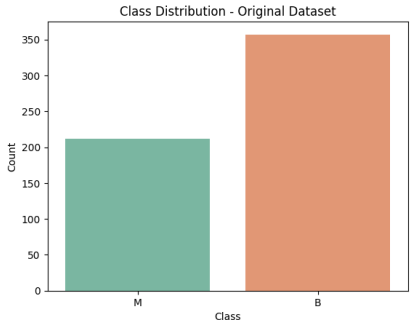



Hình 2.3: Test 60%

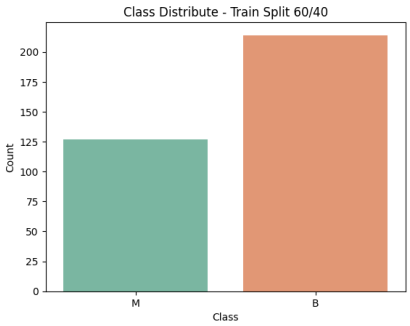
Hình 2.4: Phân phối lớp cho tỷ lệ chia 40/60.

### Tỷ lệ 60/40

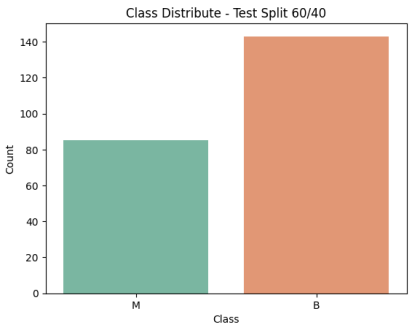
- **Tập huấn luyện**: 341 mẫu (214 Benign, 127 Malignant).
- **Tập kiểm tra**: 228 mẫu (143 Benign, 85 Malignant).




Hình 2.5: Original Dataset




Hình 2.6: Training 60%

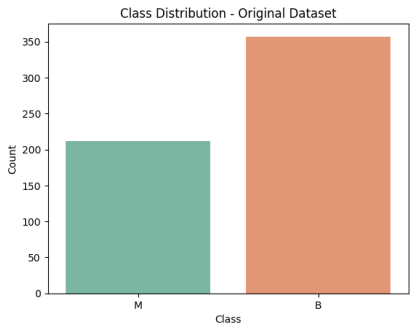



Hình 2.7: Test 40%

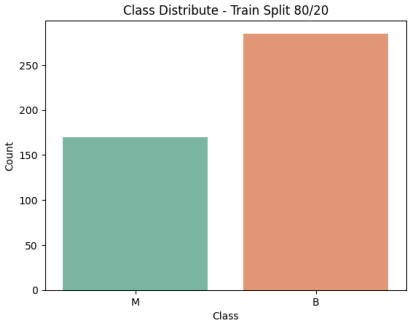
Hình 2.8: Phân phối lớp cho tỷ lệ chia 60/40.

### Tỷ lệ 80/20

- **Tập huấn luyện**: 455 mẫu (285 Benign, 170 Malignant).
- **Tập kiểm tra**: 114 mẫu (72 Benign, 42 Malignant).




Hình 2.9: Original Dataset



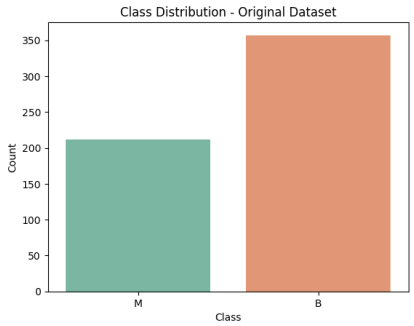

Hình 2.10: Training 80%


Hình 2.11: Test 20%

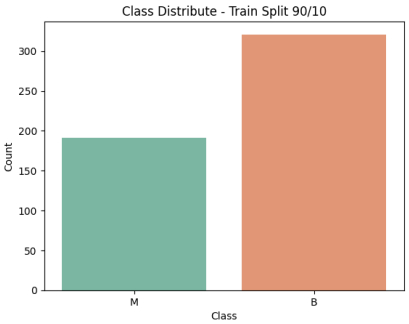
Hình 2.12: Phân phối lớp cho tỷ lệ chia 80/20.

### Tỷ lệ 90/10

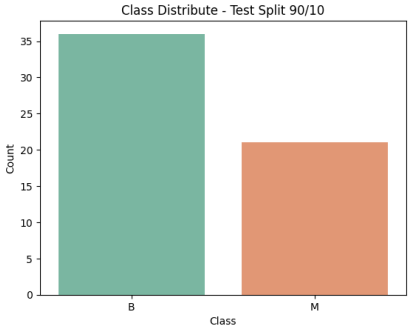
- **Tập huấn luyện**: 512 mẫu (321 Benign, 191 Malignant).
- **Tập kiểm tra**: 57 mẫu (36 Benign, 21 Malignant).




Hình 2.13: Original Dataset




Hình 2.14: Training 90%




Hình 2.15: Test 10%

Hình 2.16: Phân phối lớp cho tỷ lệ chia 90/10.

## 2.3 Phân tích các độ đo

### Classification Report và Confusion Matrix (Train=40%,Test=60%)

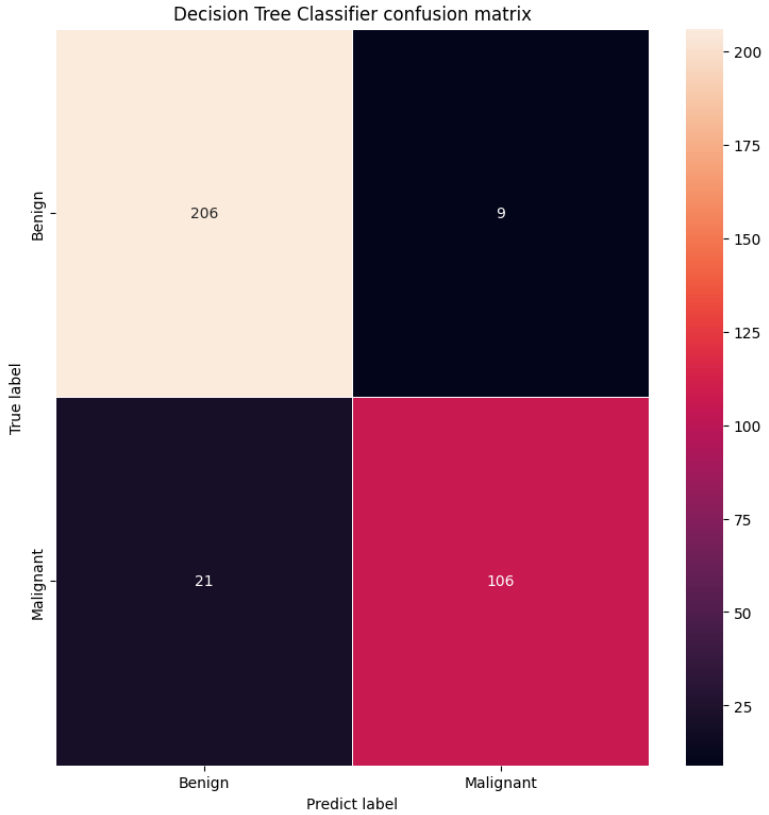
#### Phân tích Classification Report

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **B** | 0.91 | 0.96 | 0.93 | 215 |
| **M** | 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 |

Bảng 1: Classification Report với Train=40% và Test=60%.

- **Precision(Độ chính xác)** đo lường tỷ lệ dự đoán đúng trên tổng số dự đoán dương tính.
  - **B**: **0.91** cho thấy mô hình dự đoán đúng 91% các trường hợp được gán nhãn **B**.
  - **M**: **0.92** độ chính xác cao hơn **B** một chút cho thấy mô hình dự đoán đúng 92% các trường hợp được gán nhãn **M**.
  - **Trung bình (Macro Avg và Weighted Avg)**: 0.91 cho thấy kết quả ổn định và cân bằng giữa các nhãn **B** và **M**.
- **Recall(Độ nhạy)** đo lường khả năng mô hình phát hiện hết các trường hợp dương tính thực tế.
  - **B**: **0.96** cho thấy mô hình dự đoán đúng đến 96% các trường hợp thực tế được gán nhãn **B**.
  - **M**: **0.83** thấp hơn **B** cho thấy mô hình bỏ sót 17% các trường hợp thực tế thuộc nhãn **M** .Có thể là do số lượng mẫu nhãn **M** ít hơn (127 so với 215 của nhãn **B**).
  - **Trung bình (Macro Avg): 0.90, Weighted Avg: 0.91** cho thấy kết quả ổn định và cân bằng giữa các nhãn **B** và **M**.
- **F1-Score** là trung bình điều hòa giữa Precision và Recall, giúp cân bằng giữa hai chỉ số này.
  - **B**: **0.93** cho thấy mô hình đạt hiệu suất tốt cho nhãn này.
  - **M**: **0.88** thấp hơn **B** phản ánh sự cân bằng giữa Precision cao và Recall thấp.
  - **Accuracy**: 0.91 cho thấy mô hình dự đoán chính xác 91% tổng 342 mẫu.
  - **Trung bình (Macro Avg): 0.90, Weighted Avg: 0.91** cho thấy hiệu suất tổng thể tốt và cân bằng,mô hình không quá thiên lệch về một nhãn nào.

### Phân tích Confusion Matrix




Hình 2.17: Confusion Matrix với Train=40% và Test=60%

$$\begin{bmatrix} \text{TP}_B = 206 & \text{FP}_B = 21 \\ \text{FN}_B = 9 & \text{TP}_M = 106 \end{bmatrix}$$

- **True Positives (TP):**
  - **Lớp B:** 206 mẫu B được dự đoán đúng là B.
  - **Lớp M:** 106 mẫu M được dự đoán đúng là M.
- **False Positives (FP):**
  - **Lớp B:** 21 mẫu M bị dự đoán nhầm thành B.
  - **Lớp M:** 9 mẫu B bị dự đoán nhầm thành M.
- **False Negatives (FN):**
  - **Lớp B:** 9 mẫu B bị nhầm thành M.
  - **Lớp M:** 21 mẫu M bị nhầm thành B.

**Nhận xét:**

- **Hiệu suất mô hình:**

  - Mô hình hoạt động **tốt trên nhãn B** với số lượng **True Positives (TP)** rất cao: 206/215.
  - Chỉ có 9 **False Negatives (FN)** cho nhãn B → Recall của nhãn B rất cao (0.96), cho thấy mô hình ít bỏ sót các trường hợp thuộc nhãn này.
  - Đối với nhãn M, mô hình có 106 **True Positives (TP)** nhưng có 21 **False Negatives (FN)** → Recall của nhãn M thấp hơn (0.83).

- **Mất cân bằng dữ liệu:**
  - Số lượng mẫu của nhãn B (**215**) lớn hơn nhãn M (**127**), tạo ra sự **mất cân bằng** trong dữ liệu.
  - Điều này khiến mô hình có xu hướng ưu tiên dự đoán nhãn B hơn nhãn M, dẫn đến Recall thấp cho nhãn M.

- **Tổng kết hiệu suất mô hình:**
  - Mô hình đạt **Accuracy** = 0.91 → hiệu suất tổng thể rất tốt.
  - Tuy nhiên, hiệu suất giữa hai nhãn **không cân bằng**:
    * **Nhãn B:** Hoạt động xuất sắc với Recall và Precision cao.
    * **Nhãn M:** Còn hạn chế với Recall thấp do số lượng **False Negatives (21)** lớn.

**Classification Report và Confusion Matrix (Train=60%,Test=40%)**

**Phân tích Classification Report**

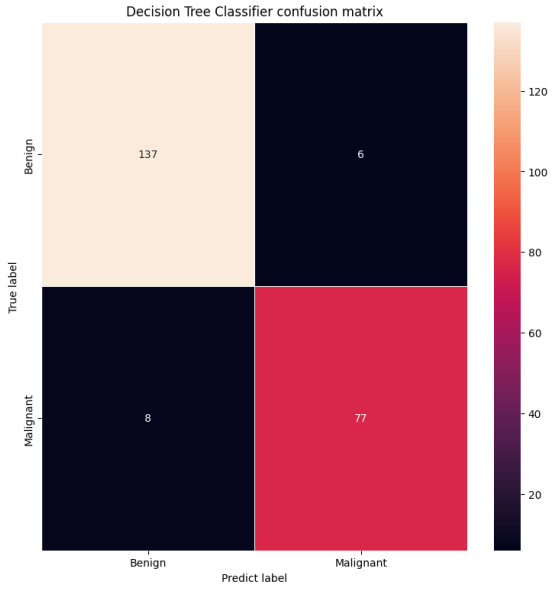
| | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| **B** | 0.94 | 0.96 | 0.95 | 143 |
| **M** | 0.93 | 0.91 | 0.92 | 85 |
| **Accuracy** | | | 0.94 | 228 |
| **Macro Avg** | 0.94 | 0.93 | 0.93 | 228 |
| **Weighted Avg** | 0.94 | 0.94 | 0.94 | 228 |

Bảng 2: Classification Report với Train=60% và Test=40%.

- **Precision (Độ chính xác)**: Precision đo lường tỷ lệ các dự đoán dương tính đúng trên tổng số dự đoán dương tính. Điều này phản ánh mức độ chính xác khi mô hình khẳng định một mẫu thuộc về một nhãn cụ thể.
  - **B (0.94):** Precision cao cho thấy mô hình rất ít dự đoán nhầm các mẫu khác thành nhãn B.
  - **M (0.93):** Precision của nhãn M đạt 93%, chỉ thấp hơn một chút so với nhãn B. Điều này chứng tỏ mô hình có khả năng xác định chính xác nhãn M, dù số lượng mẫu của nhãn này ít hơn.
  - **Trung bình (Macro Avg và Weighted Avg)**: **0.94**, phản ánh kết quả ổn định và cân bằng giữa các nhãn **B** và **M**.

- **Recall (Độ nhạy)**: Recall đo lường khả năng mô hình phát hiện đầy đủ các trường hợp dương tính thực tế. Chỉ số này đặc biệt quan trọng với các bài toán mà bỏ sót nhãn dương tính có thể gây ra rủi ro lớn.
  - **B (0.96):** Với Recall đạt 96%, mô hình phát hiện gần như tất cả các mẫu thực tế thuộc nhãn B. Điều này cho thấy mô hình rất nhạy trong việc nhận diện nhãn B và ít bỏ sót.
  - **M (0.91):** Recall của nhãn M đạt 91%, thấp hơn nhãn B. Điều này cho thấy mô hình còn bỏ sót khoảng 9% các trường hợp thực tế thuộc nhãn M, có thể do nhãn này có ít dữ liệu hơn.
  - **Trung bình (Macro Avg)**: **0.93** và **Weighted Avg**: **0.94** → mô hình hoạt động ổn định trên cả hai nhãn.
- **F1-Score**: F1-Score là trung bình điều hòa giữa Precision và Recall, giúp cân bằng giữa độ chính xác và độ nhạy.
  - **B (0.95):** F1-Score của nhãn B đạt 0.95, phản ánh khả năng cân bằng rất tốt giữa Precision và Recall. Mô hình hoạt động hiệu quả và ít mắc sai sót với nhãn này.
  - **M (0.92):** F1-Score của nhãn M đạt 0.92, thấp hơn một chút so với nhãn B do Recall thấp hơn. Mặc dù vậy, mô hình vẫn duy trì hiệu suất tốt cho nhãn này.
  - **Trung bình (Macro Avg và Weighted Avg)**: **0.93 – 0.94** cho thấy hiệu suất tổng thể tốt và mô hình không bị thiên lệch về một nhãn nào.
- **Accuracy (Độ chính xác tổng thể)**: Accuracy đạt **0.94**, nghĩa là mô hình dự đoán chính xác 94% tổng số 228 mẫu. Đây là kết quả rất tốt và phản ánh độ tin cậy cao của mô hình trên cả hai nhãn.

**Phân tích Confusion Matrix**




Hình 2.18: Confusion Matrix với Train=60% và Test=40%

$$\begin{bmatrix} \text{TP}_B = 137 & \text{FP}_B = 8 \\ \text{FN}_B = 6 & \text{TP}_M = 77 \end{bmatrix}$$

- **True Positives (TP):**
  - **Lớp B:** 137 mẫu B được dự đoán đúng là B.
  - **Lớp M:** 77 mẫu M được dự đoán đúng là M.
- **False Positives (FP):**
  - **Lớp B:** 8 mẫu M bị dự đoán nhầm thành B.
  - **Lớp M:** 6 mẫu B bị dự đoán nhầm thành M.
- **False Negatives (FN):**
  - **Lớp B:** 6 mẫu B bị nhầm thành M.
  - **Lớp M:** 8 mẫu M bị nhầm thành B.

**Nhận xét:**

- **Hiệu suất mô hình:**
  - Mô hình hoạt động **tốt trên nhãn B** với số lượng **True Positives (TP)** rất cao: 137/143.
  - Chỉ có 6 **False Negatives (FN)** cho nhãn B → Recall của nhãn B rất cao (0.96).
  - Đối với nhãn M, mô hình có 77 **True Positives (TP)** nhưng có 8 **False Negatives (FN)** → Recall của nhãn M thấp hơn một chút (0.91).
- **Sự cân bằng giữa Precision và Recall:**
  - Precision và Recall của nhãn B đều cao, chứng tỏ mô hình nhận diện tốt nhãn này.
  - Precision và Recall của nhãn M vẫn khá tốt nhưng còn một số sai sót (FN=8 và FP=8), có thể cải thiện thêm.
- **Tổng kết hiệu suất mô hình:**
  - Mô hình đạt kết quả tốt với độ chính xác cao, đặc biệt trên nhãn B.
  - Nhãn M hoạt động ổn định nhưng cần giảm tỷ lệ nhầm lẫn để tăng Recall và Precision.

**Classification Report và Confusion Matrix (Train=80%,Test=20%)**

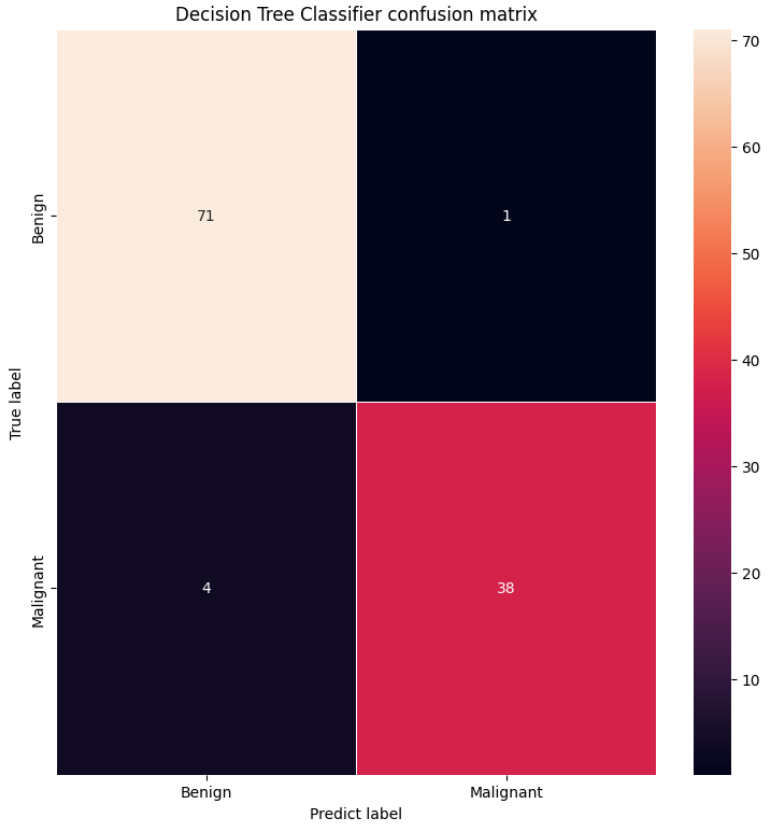
**Phân tích Classification Report**

| | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| **B** | 0.95 | 0.99 | 0.97 | 72 |
| **M** | 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 |

Bảng 3: Classification Report với Train=80% và Test=20%.

- **Precision (Độ chính xác)**: Precision đo lường tỷ lệ dự đoán dương tính đúng trên tổng số dự đoán dương tính.
  - **B (0.95):** Precision của nhãn B đạt 95%, cho thấy mô hình rất ít nhầm lẫn khi dự đoán nhãn B.
  - **M (0.97):** Precision của nhãn M cao hơn nhãn B (97%), chứng tỏ mô hình dự đoán rất chính xác cho nhãn M.
  - **Trung bình (Macro Avg và Weighted Avg)**: **0.96** cho thấy kết quả chính xác và ổn định giữa các nhãn.
- **Recall (Độ nhạy)**: Recall đo lường khả năng mô hình phát hiện đầy đủ các trường hợp dương tính thực tế.
  - **B (0.99):** Recall rất cao (99%) cho thấy mô hình phát hiện gần như toàn bộ các mẫu thực tế thuộc nhãn B.
  - **M (0.90):** Recall của nhãn M chỉ đạt 90%, thấp hơn nhãn B. Điều này cho thấy mô hình bỏ sót khoảng 10% các trường hợp thực tế thuộc nhãn M.
  - **Trung bình (Macro Avg): 0.95** và **Weighted Avg: 0.96**, thể hiện mô hình có khả năng phát hiện tốt trên cả hai nhãn.
- **F1-Score**: F1-Score là trung bình điều hòa giữa Precision và Recall, giúp cân bằng giữa độ chính xác và độ nhạy.
  - **B (0.97):** F1-Score của nhãn B đạt 0.97, phản ánh hiệu suất rất tốt nhờ Recall cực cao và Precision ổn định.
  - **M (0.94):** F1-Score của nhãn M thấp hơn một chút (**0.94**) do Recall kém hơn so với Precision.
  - **Trung bình (Macro Avg): 0.95** và **Weighted Avg: 0.96** cho thấy mô hình đạt hiệu suất tổng thể cao.
- **Accuracy (Độ chính xác tổng thể)**: Accuracy đạt **0.96**, nghĩa là mô hình dự đoán chính xác 96% trên tổng số **114** mẫu. Đây là kết quả rất tốt, cho thấy mô hình hoạt động ổn định trên cả hai nhãn.

### Phân tích Confusion Matrix




Hình 2.19: Confusion Matrix với Train=80% và Test=20%

$$\begin{bmatrix} \text{TP}_B = 71 & \text{FP}_B = 4 \\ \text{FN}_B = 1 & \text{TP}_M = 38 \end{bmatrix}$$

- **True Positives (TP):**
  - **Lớp B:** 71 mẫu B được dự đoán đúng là B.
  - **Lớp M:** 38 mẫu M được dự đoán đúng là M.
- **False Positives (FP):**
  - **Lớp B:** 4 mẫu M bị dự đoán nhầm thành B.
  - **Lớp M:** 1 mẫu B bị dự đoán nhầm thành M.
- **False Negatives (FN):**
  - **Lớp B:** 1 mẫu B bị nhầm thành M.
  - **Lớp M:** 4 mẫu M bị nhầm thành B.

**Nhận xét:**

- **Hiệu suất mô hình:**

  - **Nhãn B:** Mô hình hoạt động rất tốt với **True Positives (TP)** cao (**71/72**), chỉ có 1 **False Negative (FN)** → Recall đạt **0.99**.
  - **Nhãn M:** Mô hình có **38 True Positives (TP)** nhưng vẫn còn 4 **False Negatives (FN)**, khiến Recall đạt **0.90**.

- **Sự cân bằng Precision và Recall:**
  - Precision của nhãn B rất cao do số lượng False Positives (FP) nhỏ.
  - Recall của nhãn M (**0.90**) tuy tốt nhưng thấp hơn nhãn B (**0.99**) do số lượng **False Negatives (FN)** còn tồn tại.

- **Tổng kết hiệu suất mô hình:**
  - Độ chính xác tổng thể của mô hình rất cao với số lượng nhầm lẫn rất nhỏ.
  - **Accuracy cao và hiệu suất tốt trên cả hai nhãn**, đặc biệt là nhãn B.
  - Mô hình có thể cải thiện Recall của nhãn M bằng cách giảm **False Negatives**.

**Classification Report và Confusion Matrix (Train=90%,Test=10%)**

**Phân tích Classification Report**

| | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| **B** | 0.95 | 0.97 | 0.96 | 36 |
| **M** | 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 |

Bảng 4: Classification Report với Train=90% và Test=10%.

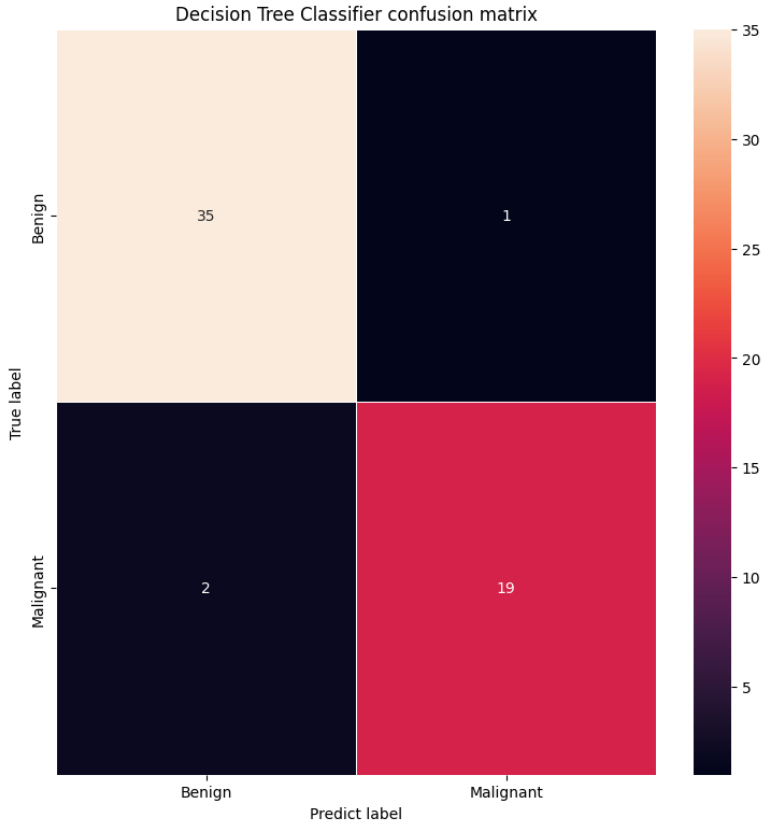
- **Precision (Độ chính xác)**: Precision đo tỷ lệ các dự đoán dương tính đúng trên tổng số dự đoán dương tính.
  - **B (0.95):** Precision của nhãn B đạt **95%**, cho thấy mô hình dự đoán chính xác hầu hết các trường hợp thuộc nhãn này.
  - **M (0.95):** Precision của nhãn M cũng đạt **95%**, phản ánh độ chính xác tương đương với nhãn B.
  - **Trung bình (Macro Avg và Weighted Avg):** Cả hai giá trị đều là **0.95**, chứng tỏ mô hình dự đoán ổn định và cân bằng giữa các nhãn.

- **Recall (Độ nhạy)**: Recall phản ánh khả năng phát hiện đầy đủ các trường hợp dương tính thực tế.
  - **B (0.97):** Recall của nhãn B đạt **97%**, nghĩa là mô hình gần như phát hiện toàn bộ các mẫu thực tế thuộc nhãn B.
  - **M (0.90):** Recall của nhãn M thấp hơn (**90%**), cho thấy khoảng **10%** các trường hợp nhãn M bị bỏ sót.

  - **Trung bình (Macro Avg): 0.94** cho thấy mức độ phát hiện tổng thể trên cả hai nhãn vẫn tốt nhưng bị ảnh hưởng bởi Recall của nhãn M.

- **F1-Score**: F1-Score là trung bình điều hòa giữa Precision và Recall, giúp cân bằng độ chính xác và độ nhạy.
  - **B (0.96):** F1-Score của nhãn B đạt **0.96**, phản ánh hiệu suất rất cao với độ chính xác và độ nhạy tốt.
  - **M (0.93):** F1-Score của nhãn M đạt **0.93**, thấp hơn nhãn B do Recall còn hạn chế.
  - **Trung bình (Macro Avg): 0.94** và **Weighted Avg: 0.95**, chứng tỏ mô hình hoạt động hiệu quả trên toàn bộ tập dữ liệu.
- **Accuracy (Độ chính xác tổng thể)**: Accuracy đạt **0.95**, nghĩa là mô hình dự đoán chính xác 95% trên tổng số **57** mẫu. Đây là kết quả rất tốt, phản ánh khả năng tổng quát hóa tốt của mô hình trên tập dữ liệu test.

### Phân tích Confusion Matrix




Hình 2.20: Confusion Matrix với Train=90% và Test=10%

$$\begin{bmatrix} \text{TP}_B = 35 & \text{FP}_B = 2 \\ \text{FN}_B = 1 & \text{TP}_M = 19 \end{bmatrix}$$

- **True Positives (TP):**

  - **Lớp B:** 35 mẫu B được dự đoán đúng là B.
  - **Lớp M:** 19 mẫu M được dự đoán đúng là M.

- **False Positives (FP):**
  - **Lớp B:** 2 mẫu M bị dự đoán nhầm thành B.
  - **Lớp M:** 1 mẫu B bị dự đoán nhầm thành M.

- **False Negatives (FN):**
  - **Lớp B:** 1 mẫu B bị nhầm thành M.
  - **Lớp M:** 2 mẫu M bị nhầm thành B.

**Nhận xét:**

- **Hiệu suất mô hình:**
  - **Nhãn B:** Mô hình hoạt động rất tốt với **True Positives (TP)** đạt **35/36**, chỉ có 1 **False Negative (FN)**. Điều này giúp Recall của nhãn B đạt **0.97**.
  - **Nhãn M:** Mô hình có **19 True Positives (TP)** nhưng còn 2 **False Negatives (FN)**, làm Recall của nhãn M giảm xuống **0.90**.

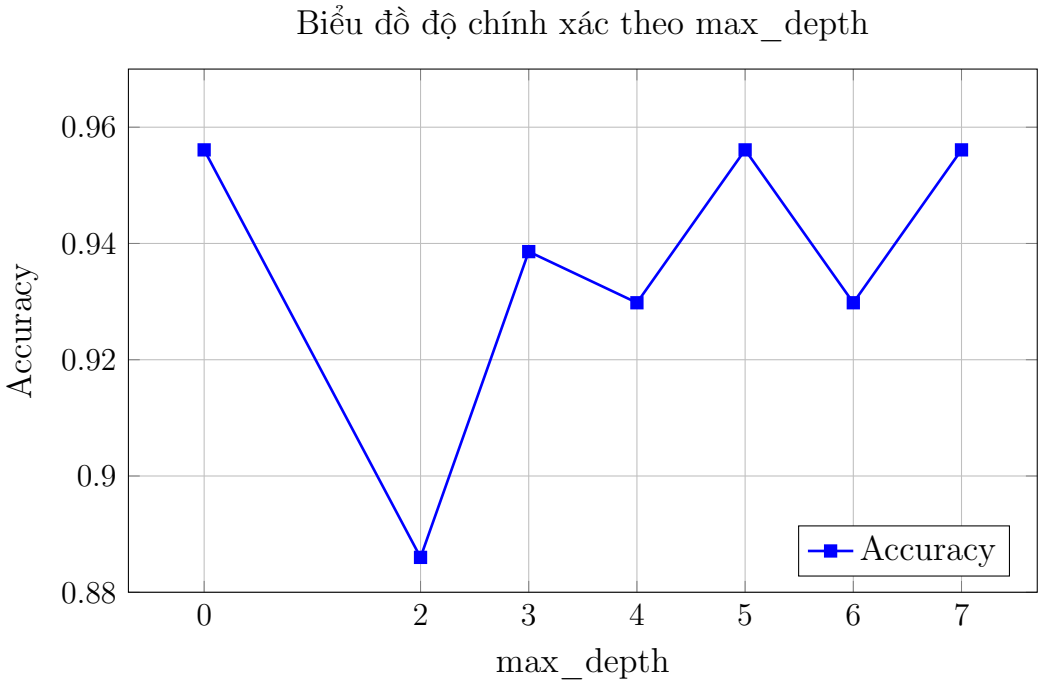
- **Sự cân bằng Precision và Recall:**
  - Precision của nhãn B rất cao do số lượng False Positives (FP) nhỏ (**2**).
  - Precision và Recall của nhãn M còn cải thiện được bằng cách giảm False Negatives (**2**) và False Positives (**1**).

- **Tổng kết hiệu suất mô hình:**
  - Độ chính xác tổng thể của mô hình rất cao với mức nhầm lẫn rất nhỏ.
  - Nhãn B có độ chính xác và độ nhạy cao vượt trội.
  - Nhãn M cũng có hiệu suất tốt nhưng cần tối ưu để giảm thiểu số lượng False Negatives.

## 2.4 Phân tích độ sâu và độ chính xác

| **max_depth** | **None** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| **Accuracy** | 0.9561 | 0.8860 | 0.9386 | 0.9298 | 0.9561 | 0.9298 | 0.9561 |

Bảng 5: Độ chính xác với từng giá trị max_depth.




Hình 2.21: Biểu đồ độ chính xác với từng giá trị max_depth.

**Nhận xét:**

- **max_depth = None:** Độ chính xác đạt **0.9561**, rất cao. Không giới hạn độ sâu giúp mô hình phân tách hoàn hảo nhưng có thể dẫn đến quá khớp (overfitting).
- **max_depth = 2:** Độ chính xác giảm xuống **0.8860**. Điều này cho thấy mô hình quá đơn giản, không đủ độ sâu để phân tách dữ liệu chính xác.
- **max_depth = 3:** Độ chính xác tăng lên **0.9386**. Mô hình cải thiện so với ‘max_depth = 2‘ nhưng vẫn chưa đạt hiệu suất tối đa.
- **max_depth = 4:** Độ chính xác giảm nhẹ còn **0.9298**, sự cân bằng giữa underfitting và overfitting chưa được tối ưu.
- **max_depth = 5:** Độ chính xác quay lại mức cao nhất là **0.9561**, cho thấy đây có thể là độ sâu tối ưu cho mô hình này.
- **max_depth = 6:** Độ chính xác giảm nhẹ xuống **0.9298**, cho thấy mô hình có thể đã bắt đầu xuất hiện dấu hiệu của quá khớp.
- **max_depth = 7:** Độ chính xác đạt lại **0.9561**, giống với ‘max_depth = 5‘. Mức này đảm bảo mô hình hoạt động tốt nhưng cần kiểm tra thêm để tránh quá khớp(overfitting).

# 3 Tập dữ liệu thứ hai

## 3.1 Tiền xử lý dữ liệu

## 3.2 Xây dựng các cây quyết định

## 3.3 Phân tích các độ đo

# 4 Tập dữ liệu thứ ba

## 4.1 Tiền xử lý dữ liệu

## 4.2 Xây dựng các cây quyết định

## 4.3 Phân tích các độ đo

# 5 Phân tích độ sâu của cây quyết định

## 5.1 Phân tích và so sánh 3 bộ dữ liệu

**Sơ lược về cả 3 bộ dữ liệu**

**Breast Cancer Wisconsin (Diagnostic)**

- **Loại bài toán:** Binary Classification .
- **Đặc trưng:** 30 thuộc tính.
- **Kích thước:** 569 mẫu.
- **Mục tiêu:** Dự đoán khối u là lành tính (Benign) hay ác tính (Malignant).

**Wine Quality**

- **Loại bài toán:** Multiclass Classification.
- **Đặc trưng:** 11 thuộc tính.
- **Kích thước:** 4,898 mẫu.
- **Mục tiêu:** Dự đoán chất lượng rượu trong 3 nhóm: High, Standard, Low.

**....**

- **Loại bài toán:**
- **Đặc trưng:**
- **Kích thước:**
- **Mục tiêu:**

**So sánh dựa trên Classification Report**

**Hiệu suất chung của mô hình**

- **Breast Cancer Wisconsin (Diagnostic):**
  - Độ chính xác cao và ổn định, dao động từ **91% đến 96%** khi tăng tỷ lệ huấn luyện.
  - Macro avg (**Precision**, **Recall**, **F1-Score**) luôn trên **0.90**.
- **Wine Quality:**
  - Độ chính xác thấp hơn, chỉ dao động từ **74% đến 79%**.
  - Macro avg thấp hơn, chỉ đạt khoảng **0.52 – 0.62** do ảnh hưởng từ mất cân bằng lớp.

**Ảnh hưởng của class imbalance**

- **Breast Cancer Wisconsin (Diagnostic):**
  - Hiệu suất tăng rõ rệt khi tăng tỷ lệ huấn luyện → mô hình học tốt hơn khi có nhiều dữ liệu.
- **Wine Quality:**
  - Hiệu suất tăng nhẹ khi tăng tỷ lệ huấn luyện, nhưng không đáng kể → mô hình khó khái quát hóa trên bài toán đa lớp.

**Số lớp của dữ liệu**

- **Breast Cancer Wisconsin (Diagnostic):**
  - Bài toán nhị phân (Binary Classification): đơn giản, ít lớp (B hoặc M) giúp mô hình phân tách hiệu quả và tính toán nhanh.
- **Wine Quality:**
  - Bài toán phân loại đa lớp (Multi-Class Classification): khó hơn do phải xử lý 3 lớp (High, Standard, Low), tăng độ phức tạp cho mô hình.

**Số thuộc tính (features)**

- **Breast Cancer Wisconsin (Diagnostic):**
  - Có **30 thuộc tính**, đủ lớn để mô hình học tốt nhưng không gây overfitting khi tăng dữ liệu huấn luyện.
- **Wine Quality:**
  - Chỉ có **11 thuộc tính**, tuy nhỏ hơn nhưng dữ liệu đa lớp và mất cân bằng khiến hiệu suất bị giới hạn.

**Kích thước tập dữ liệu**

- **Breast Cancer Wisconsin (Diagnostic):**
  - Dữ liệu nhỏ (**569 mẫu**), nhưng do bài toán đơn giản, mô hình vẫn duy trì hiệu suất cao.
- **Wine Quality:**
  - Dữ liệu lớn hơn (**4,898 mẫu**), cung cấp nhiều thông tin nhưng không cải thiện đáng kể hiệu suất do bài toán khó khái quát hóa.

**So sánh dựa trên max_depth**

**Breast Cancer Wisconsin (Diagnostic)**

- **Hiệu suất mô hình:** Tập dữ liệu này có độ chính xác rất cao, ngay cả khi độ sâu của cây quyết định ở mức trung bình (**5–7**).Cho thấy:
  - Mô hình học tốt các đặc trưng của dữ liệu.
  - Tính chất của bài toán phân loại nhị phân giúp việc phân tách dữ liệu dễ dàng hơn.
- **Số lượng đặc trưng và kích thước dữ liệu:**
  - **Số lượng đặc trưng lớn:** Với **30 thuộc tính**, mô hình cây quyết định có nhiều điều kiện phân tách để lựa chọn, giúp tăng khả năng phân loại chính xác.
  - **Kích thước dữ liệu nhỏ:** Với chỉ **569 mẫu**, mô hình có nguy cơ bị **overfitting** khi độ sâu của cây tăng quá lớn (ví dụ: max_depth $> 7$). Điều này xảy ra do mô hình có thể học quá chi tiết các đặc trưng của dữ liệu huấn luyện, dẫn đến giảm khả năng khái quát hóa trên tập dữ liệu mới.
- **Kiểm soát overfitting:** Việc chọn giá trị **max_depth** thích hợp (từ **5–7**) là rất quan trọng để cân bằng giữa:
  - **Hiệu suất cao:** Độ chính xác tốt nhất đạt **0.9561**.
  - **Tránh overfitting:** Độ sâu lớn có thể khiến mô hình quá khớp với dữ liệu huấn luyện.

**Wine Quality**

- **Độ phức tạp của bài toán:** Bài toán phân loại đa lớp (ví dụ: **Wine Quality**) có tính chất phức tạp hơn so với bài toán phân loại nhị phân, dẫn đến:
  - Mô hình cần phải xử lý nhiều điều kiện phân tách hơn để phân biệt giữa các lớp.
  - Độ chính xác của mô hình thường thấp hơn, do khó khăn trong việc học và khái quát hóa mối quan hệ phức tạp giữa các đặc trưng và nhãn lớp.
- **Tầm quan trọng của việc chọn giá trị max_depth:** Để mô hình hoạt động hiệu quả, việc chọn giá trị **max_depth** tối ưu là rất quan trọng:

- **Học đủ mối quan hệ phức tạp:** Độ sâu phù hợp giúp mô hình học tốt các mối quan hệ phức tạp giữa các lớp trong bài toán đa lớp.
- **Tránh overfitting:** Độ sâu quá lớn có thể dẫn đến tình trạng mô hình học quá chi tiết dữ liệu huấn luyện, làm giảm khả năng khái quát hóa trên tập dữ liệu mới.

**.....**

# Tài liệu tham khảo