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 2

MÔN CƠ SỞ TRÍ TUỆ NHÂN TẠO

LỚP: 22_21

GIÁO VIÊN HƯỚNG DẪN

NGUYỄN THANH TÌNH

TP.HỒ CHÍ MINH – NĂM 2024

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

Mục lục

1. Thông tin thành viên

STT	MSSV	Họ và tên
1	21120042	Phan Gia Bảo
2	21120097	Trần Bảo Minh
3	22120422	Nguyễn Phạm Tú Uyên
4	22120449	Lê Nguyễn Huyền Vy

2. Bảng phân công công việc

STT	Họ và tên	Phân công	Mức độ hoàn thiện
1	Phan Gia Bảo	- Tổng hợp bài - Deploy notebook sử dụng docker - Viết báo cáo	100%
2	Trần Bảo Minh	- Additional dataset - Viết báo cáo	100%
3	Nguyễn Phạm Tú Uyên	- Breast Cancer dataset - Viết báo cáo	100%
4	Lê Nguyễn Huyền Vy	- Wine Quality dataset - Viết báo cáo	100%

3. Đánh giá mức độ hoàn thành

STT	Công việc	Chi tiết	Mức độ hoàn thiện
1	Additional dataset	Cài đặt và phân tích Insight	100%
2	Breast Cancer dataset	Cài đặt và phân tích Insight	100%
3	Wine Quality dataset	Cài đặt và phân tích Insight	100%
4	Tổng hợp bài	Chỉnh sửa, review lại tất cả notebook	100%
5	Deploy notebook sử dụng docker	Deploy notebook bằng docker và viết hướng dẫn chạy dockerfile	100%
6	Viết báo cáo		100%

4. Thuật toán Cây quyết định (Decision Tree)

a. Giới thiệu

Thuật toán Decision Tree (Cây quyết định) là một trong những thuật toán phổ biến trong học máy, được sử dụng cho cả bài toán phân loại và hồi quy. Đây là một mô hình dựa trên cây, nơi mỗi nút đại diện cho một điều kiện phân tách dựa trên giá trị của một thuộc tính, các

nhánh con đại diện cho kết quả của điều kiện phân tách, và các nút lá biểu diễn kết quả đầu ra cuối cùng.

b. Một số thuật toán xây dựng Cây quyết định (Decision Tree)

i. ID3 (Iterative Dichotomiser 3)

1. Giới thiệu

ID3 là một thuật toán xây dựng Decision Tree do Ross Quinlan giới thiệu. Đây là một thuật toán đơn giản và dễ hiểu, sử dụng khái niệm Entropy và Information Gain để chọn thuộc tính phân chia dữ liệu tại mỗi bước.

2. Các khái niệm chính

- **Hàm Entropy :** xác định tính không thuần khiết của một tập các ca dữ liệu bất kỳ.

$$Entropy(S) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

Trong đó:

- S : là tập dữ liệu đang xét
- p_i : xác suất xuất hiện của lớp trong tập dữ liệu S
- n : là số lượng lớp có trong tập dữ liệu hiện tại

- **Information Gain** đo mức độ hiệu quả của một thuộc tính trong bài toán phân lớp dữ liệu. Đó chính là sự rút gọn mà ta mong đợi khi phân chia các ca dữ liệu theo thuộc tính này. Nó được tính theo công thức sau đây:

$$IG(A) = Entropy(S) - \sum_{v \epsilon Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Trong đó:

- A : thuộc tính được đánh giá
- Value(A) : tập tất cả các giá trị có thể có đối với thuộc tính A
- S_v : tập con của S mà A có giá trị là v
- $|S_v|$: số lượng mẫu trong S_v
- $|S|$: tổng số lượng mẫu trong S

3. Quy trình xây dựng Cây quyết định theo từng bước

- Bước 1 : Tính Entropy cho tập dữ liệu ban đầu S, dựa trên phân phối nhãn của tập dữ liệu.
- Bước 2 :
  - Với mỗi thuộc tính A, tính Information Gain (IG) khi phân chia dữ liệu theo thuộc tính này.
  - Chọn thuộc tính A có Information Gain (IG) cao nhất để phân chia tập dữ liệu.
- Bước 3 : Phân chia tập dữ liệu thành các tập con S_v dựa trên các giá trị v của thuộc tính A.
- Bước 4 :
  - Áp dụng lại thuật toán ID3 trên mỗi tập con S_v để tiếp tục xây dựng cây.
  - Lặp lại quá trình này cho đến khi đạt tiêu chí dừng:
    + Tập con thuần nhất (chỉ chứa một lớp).
    + Không còn thuộc tính nào để phân chia.
    + Không còn mẫu nào trong tập con.

ii. C4.5

1. Giới thiệu

C4.5 là một cải tiến của thuật toán ID3, cũng được phát triển bởi Ross Quinlan. Thuật toán này giải quyết các hạn chế của ID3 và trở thành một trong những thuật toán xây dựng Decision Tree phổ biến nhất. C4.5 được biết đến nhờ khả năng xử lý thuộc tính liên tục, dữ liệu bị thiếu, và tạo ra cây quyết định gọn gàng hơn nhờ tiêu chí Gain Ratio.

2. Các khái niệm chính

- Các khái niệm về Entropy và Information Gain vẫn giống với ID3 ngoài ra ta còn có một số khái niệm mới sau.

- **Split Information** : đo lường mức độ phân chia của dữ liệu. Có công thức như sau:

$$SplitInformation(S, A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$$

Trong đó:

- $|S_v|$: là tập con của S với A có giá trị là v_i
- $|S|$: là tổng số mẫu trong tập dữ liệu gốc
- c : là số lượng giá trị khác nhau của thuộc tính A

- **Gain Ratio** : giảm trọng số của các thuộc tính có Split Information cao, vì chúng làm phân mảnh dữ liệu. Có công thức như sau :

$$GainRatio(S, A) = \frac{IG(A)}{SplitInformation(S, A)}$$

3. Quy trình xây dựng Cây quyết định theo từng bước

- Bước 1 : Giống với ID3, tính entropy tổng thể của tập dữ liệu ban đầu dựa trên phân phối nhãn.
- Bước 2 : Đối với mỗi thuộc tính liên tục:
  - Sắp xếp các giá trị thuộc tính theo thứ tự tăng dần.
  - Xác định điểm cắt (threshold) giữa các giá trị liên tiếp.
  - Tính Information Gain cho từng điểm cắt để tìm ngưỡng phân chia tối ưu.
  - Chia thuộc tính liên tục thành hai khoảng: ≤ threshold và > threshold
- Bước 3 :
  - Tính Information Gain và Split Information cho từng thuộc tính.
  - Tính Gain Ratio dựa trên công thức đã nêu.
  - Chọn thuộc tính có Gain Ratio cao nhất để phân chia dữ liệu.
- Bước 4 :
  - Phân chia dữ liệu dựa trên thuộc tính được chọn và các giá trị của nó (hoặc điểm cắt với thuộc tính liên tục).
  - Tiếp tục lặp lại quá trình trên từng tập con.
- Bước 5 :
  - Dừng lại khi:
    + Tập con chỉ còn một lớp (Entropy = 0).
    + Không còn thuộc tính nào để phân chia.
    + Số lượng mẫu trong tập con không đủ lớn (được kiểm soát bởi tham số).
- Bước 6 :
  - Sau khi xây dựng cây xong sẽ thực hiện cắt tỉa các nhánh không cần thiết.

c. Ưu điểm và nhược điểm của Cây quyết định (Decision Tree)

i. Ưu điểm

- Có thể xử lý tốt với dữ liệu dạng số (rời rạc và liên tục)
- Mô hình dễ hiểu và dễ giải thích.
- Không cần chuẩn hóa hoặc tạo biến giả cho dữ liệu

ii. Nhược điểm

- Hay gặp vấn đề overfitting
- Không đảm bảo xây dựng được cây tối ưu
- Cấu trúc cây rất dễ bị thay đổi nếu có một thay đổi nhỏ trong dữ liệu dẫn đến kết quả sai hoàn toàn

5. Xây dựng Cây quyết định (Decision Tree) trên các tập dữ liệu thực tế bằng scikit-learn

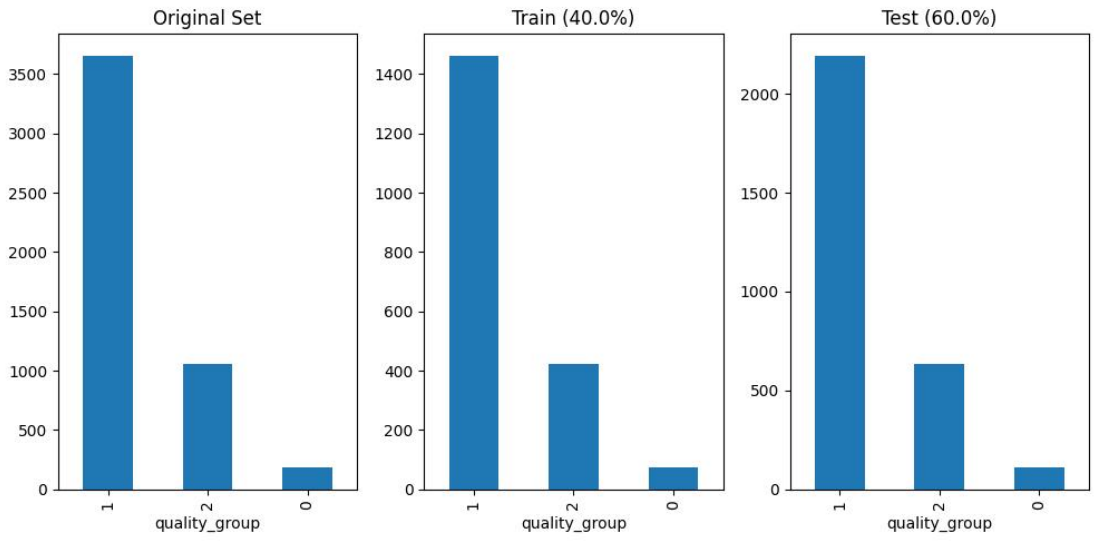
a. Wine Quality dataset

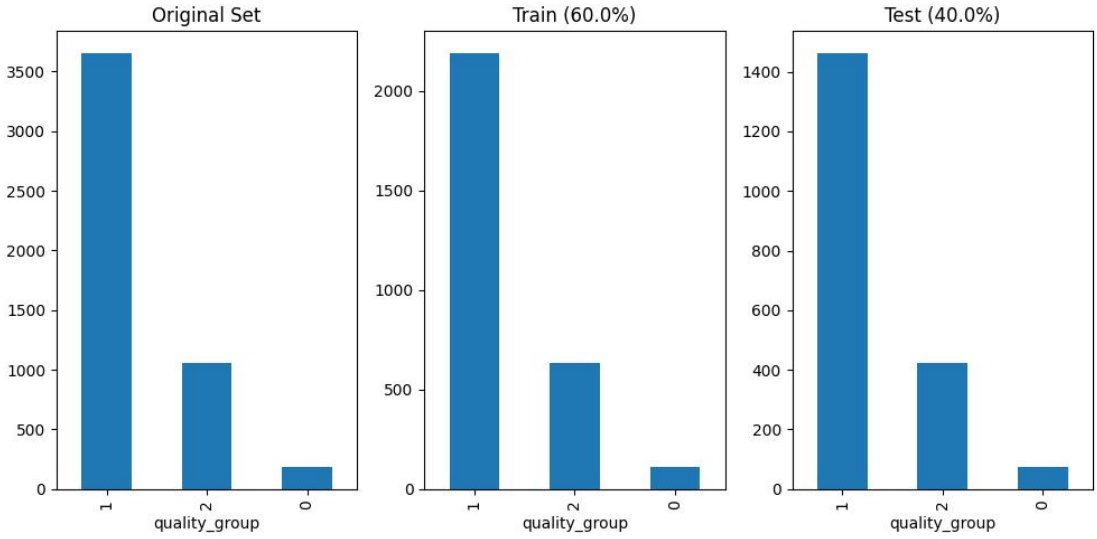
i. Giới thiệu về dữ liệu

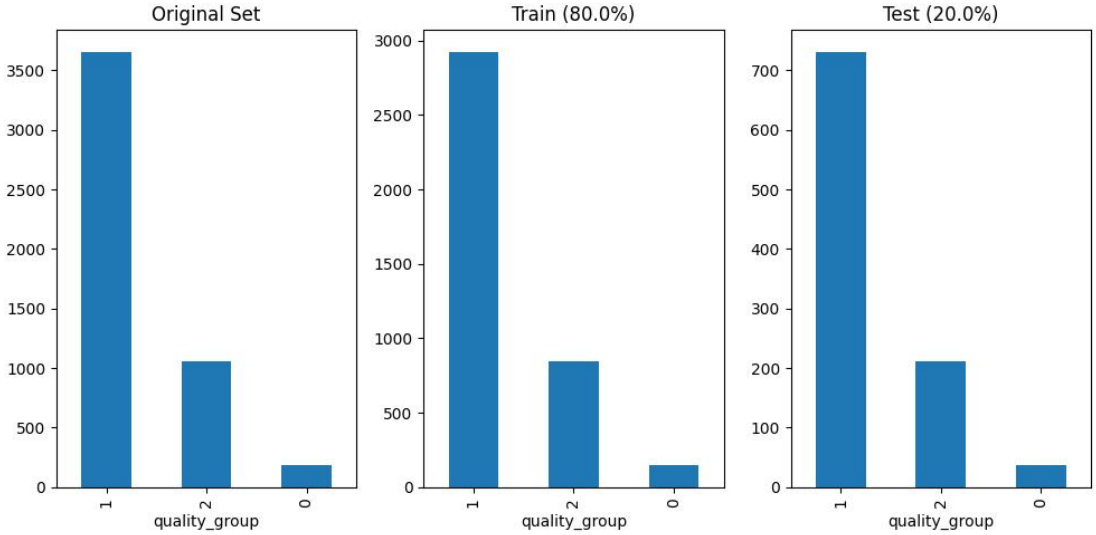
- Nguồn: The UCI Wine Quality Dataset
- Dataset gồm 4898 dòng và 11 cột, mỗi dòng mô tả về các đặc trưng hóa, lý của rượu vang.
- Ý nghĩa của các cột:
  - **fixed acidity**: độ axit cố định, không bay hơi trong rượu
  - **volatile acidity**: lượng axit bay hơi
  - **citric acid**: một loại axit tự nhiên có trong quả nho
  - **residual sugar**: lượng đường còn lại trong quá trình lên men
  - **chlorides**: hàm lượng muối clorua trong rượu
  - **free sulfur dioxide**: hàm lượng chất bảo quản (dioxide lưu huỳnh tự do) trong rượu vang.
  - **total sulfur dioxide**: tổng lượng dioxide lưu huỳnh, bao gồm cả dạng tự do và dạng kết hợp
  - **density**: (mật độ) của rượu, thường liên quan đến hàm lượng đường và cồn.
  - **pH**: chỉ số đo độ axit của rượu vang, ảnh hưởng đến vị và độ ổn định của rượu.
  - **sulphates**: lượng sunfat được thêm vào trong quá trình sản xuất.
  - **alcohol:** nồng độ cồn, phần trăm thể tích rượu trong rượu vang.
  - **quality:** điểm đánh giá chất lượng rượu.

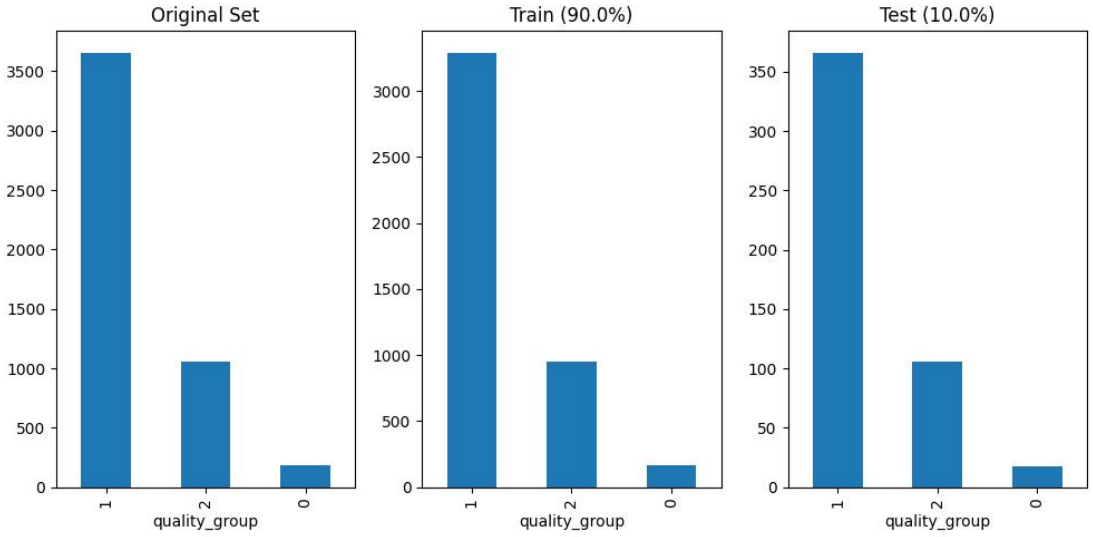
ii. Phân tích dữ liệu

1. Chuẩn bị dữ liệu (Chia dữ liệu):









- Sau khi chia dữ liệu, tỷ lệ phân bố giữa các lớp vẫn được duy trì đồng đều ở cả tập huấn luyện và kiểm tra. Đảm bảo mô hình được phân tích trên các dữ liệu có đặc điểm tương tự, giúp kết quả phân loại trở nên đáng tin cậy. Việc duy trì sự phân bố hợp lý giữa các lớp trong tất cả các tập dữ liệu chứng minh rằng việc chia dữ liệu đã được thực hiện đúng cách.

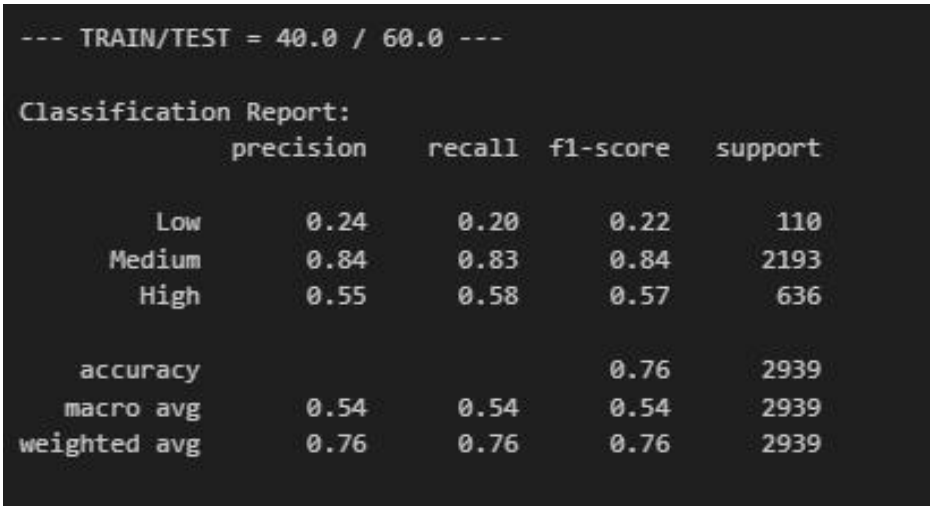
2. Xây dựng cây quyết định (Decision Tree)

Hình ảnh cây quyết định được lưu dưới định dạng file .svg trong thư mục Decision_Tree_Visualizations_WineQuantity_Dataset (thư mục có đường dẫn giống với đường dẫn đến file source code, thư mục xuất hiện sau khi chạy file WineQuality.ipynb)

3. Đánh giá kết quả qua Classification Report và Confusion Matrix.

- **Classification Report** hiển thị các chỉ số đánh giá hiệu suất mô hình trên từng lớp của dữ liệu.
  - *precision:* tỷ lệ dự đoán đúng trong số các dự đoán của mỗi lớp.
  - *recall:* tỷ lệ nhận diện đúng mẫu thuộc một lớp nào đó.
  - *f1-score:* kết hợp giữa **precision** và **recall**
  - *support: số lượng mẫu thật sự của mỗi lớp.*
  - *accuracy: độ chính xác của tổng thể mô hình*
  - *macro avg: trung bình của precision, recall và f1-score*
  - *weighted avg: trung bình dựa trên số lượng mẫu trong mỗi lớp*
- **Confusion Matrix** là ma trận nhầm lẫn thể hiện các lỗi trong dự đoán của mô hình.
  - Các phần tử nằm trên *đường chéo chính* của ma trận sẽ biểu thị số lượng các dự đoán chính xác.
  - Các phần tử còn lại là số dự đoán sai, bị nhầm từ lớp i sang lớp j. (với j là chỉ số cột, i là chỉ số dòng)

- **Trường hợp train/test = 40/60:**



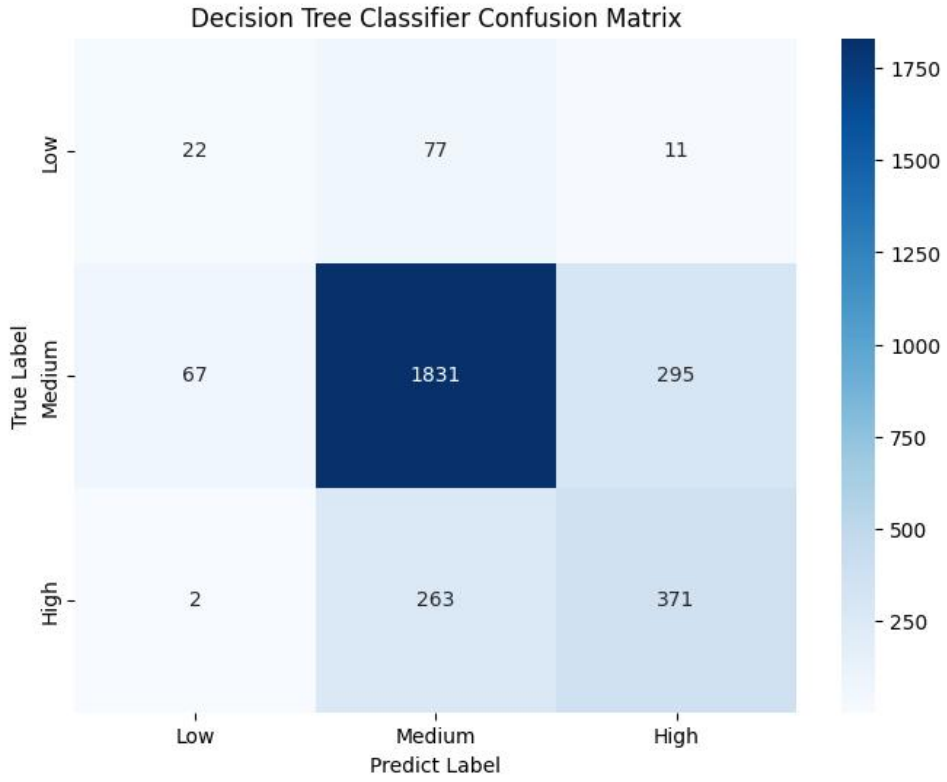
```
--- TRAIN/TEST = 40.0 / 60.0 ---

Classification Report:
              precision    recall  f1-score   support

         Low       0.24      0.20      0.22       110
      Medium       0.84      0.83      0.84      2193
        High       0.55      0.58      0.57       636

    accuracy                           0.76      2939
   macro avg       0.54      0.54      0.54      2939
weighted avg       0.76      0.76      0.76      2939
```

Trong mô hình có tổng cộng 110 mẫu thuộc lớp Low, 2193 mẫu thuộc lớp Medium, 636 mẫu thuộc lớp High, mô hình dự đoán đúng 76% số mẫu trong tập kiểm tra

  - *Lớp Low:*
    - Precision: dự đoán chính xác 24%
    - Recall: Mô hình phát hiện được 20% số mẫu thật sự thuộc lớp Low
    - F1-score: 20% cho thấy hiệu suất cân bằng của lớp Low rất thấp.
  - *Lớp Medium:*
    - Precision: dự đoán chính xác 84%
    - Recall: Mô hình phát hiện được 83% số mẫu thật sự thuộc lớp Medium
    - F1-score: 84% cho thấy hiệu suất cân bằng của lớp Medium tốt.
    - Lớp Medium trong mô hình này cũng là lớp đạt được hiệu quả cao nhất với các tỷ lệ trên 80%.

- *Lớp High:*
  - Precision: dự đoán chính xác 55%
  - Recall: Mô hình phát hiện được 58% số mẫu thật sự thuộc lớp High
  - F1-score: 57% cho thấy hiệu suất cân bằng của lớp High chỉ mới vượt qua mức trung bình.
- *Các chỉ số tính trung bình:*
  - Macro Average: trung bình giữa ba lớp: precision 0.54, recall 0.54, f1-score 0.54.
  - Weighted Average trung bình nhưng có dựa trên số mẫu của mỗi lớp trong mô hình: precision 0.76, recall 0.76, f1-score 0.76.

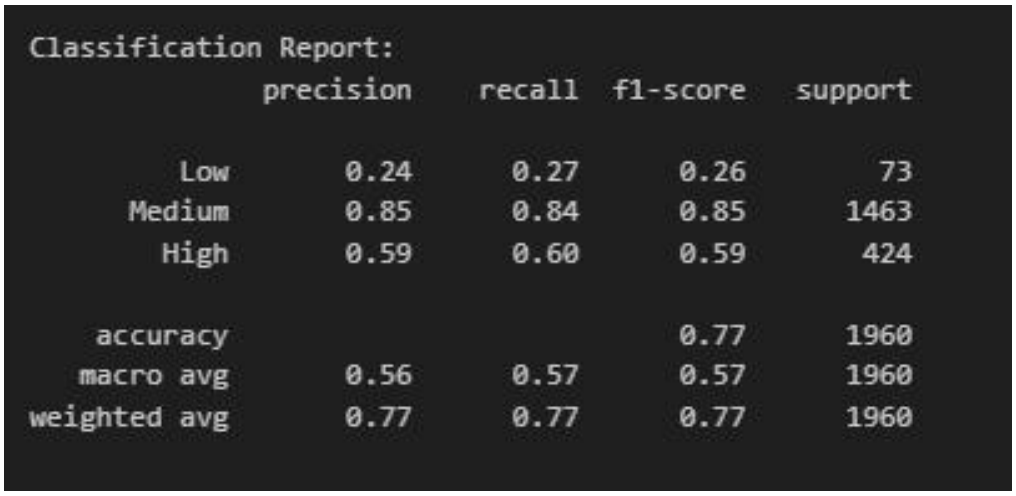
-> Có sự chênh lệch này là do lớp Low đạt độ chính xác thấp nhưng thực tế lớp Low lại có rất ít mẫu.



Nhận xét:

- Mô hình làm việc tốt nhất trong việc phân loại lớp Medium, tuy nhiên mẫu lớp Medium có 295 mẫu vẫn còn bị nhầm lẫn sang lớp High, 67 mẫu nhầm sang lớp Low.
- Và ngược lại, gần ½ mẫu từ lớp High (263 mẫu) bị nhầm lẫn sang lớp Medium.
- Lớp Low có hoạt động dự đoán kém nhất, khi có đến 88 mẫu bị nhầm sang hai lớp còn lại, chỉ có 22 mẫu được dự đoán đúng.

- **Trường hợp train/test = 60/40:**



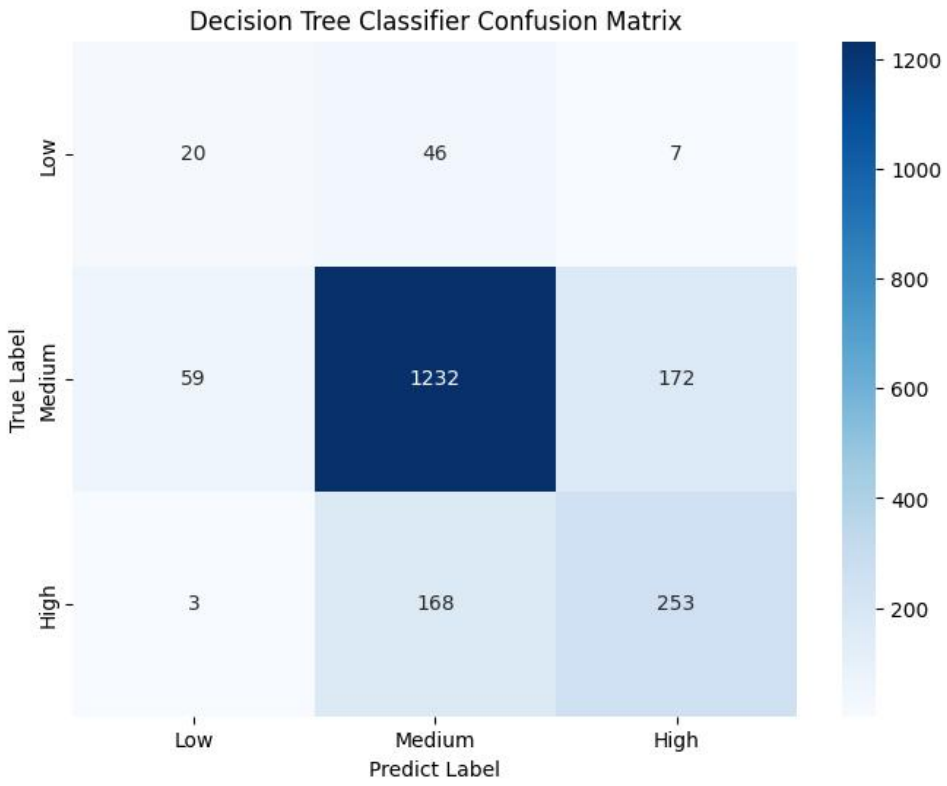
Classification Report:

	precision	recall	f1-score	support
Low	0.24	0.27	0.26	73
Medium	0.85	0.84	0.85	1463
High	0.59	0.60	0.59	424
accuracy			0.77	1960
macro avg	0.56	0.57	0.57	1960
weighted avg	0.77	0.77	0.77	1960

Trong mô hình có tổng cộng 73 mẫu thuộc lớp Low, 1463 mẫu thuộc lớp Medium, 424 mẫu thuộc lớp High, mô hình dự đoán đúng 77% số mẫu trong tập kiểm tra.

- *Lớp Low:*
  - Precision: dự đoán chính xác 24%
  - Recall: Mô hình phát hiện được 27% số mẫu thật sự thuộc lớp Low
  - F1-score: 26% cho thấy hiệu suất cân bằng của lớp Low rất thấp.
- *Lớp Medium:*
  - Precision: dự đoán chính xác 85%
  - Recall: Mô hình phát hiện được 84% số mẫu thật sự thuộc lớp Medium
  - F1-score: 85% cho thấy hiệu suất cân bằng của lớp Medium tốt.
  - Lớp Medium trong mô hình này cũng là lớp đạt được hiệu quả cao nhất với các tỷ lệ trên 80%.
- *Lớp High:*
  - Precision: dự đoán chính xác 59%
  - Recall: Mô hình phát hiện được 60% số mẫu thật sự thuộc lớp High.
  - F1-score: 59% cho thấy hiệu suất cân bằng của lớp High ở mức ổn tuy nhiên đây vẫn chưa phải một kết quả lý tưởng.
- *Các chỉ số tính trung bình:*
  - Macro Average: trung bình giữa ba lớp: precision 0.56, recall 0.57, f1-score 0.57.
  - Weighted Average trung bình nhưng có dựa trên số mẫu của mỗi lớp trong mô hình: precision 0.77, recall 0.77, f1-score 0.77.

-> Có sự chênh lệch này là do lớp Low đạt độ chính xác thấp nhưng thực tế lớp Low lại có rất ít mẫu.



Nhận xét:

- Mô hình làm việc tốt nhất trong việc phân loại lớp Medium, tuy nhiên mẫu lớp Medium đa số vẫn còn bị nhầm lẫn sang lớp High, có 172 mẫu bị nhầm sang lớp High, và 59 mẫu bị nhầm sang lớp Low.
- Và ngược lại, gần ⅓ mẫu từ lớp High (168 mẫu) bị nhầm lẫn sang lớp Medium.
- Lớp Low có hoạt động dự đoán kém nhất, khi hơn phân nửa số mẫu của lớp Low bị nhầm sang hai lớp còn lại.

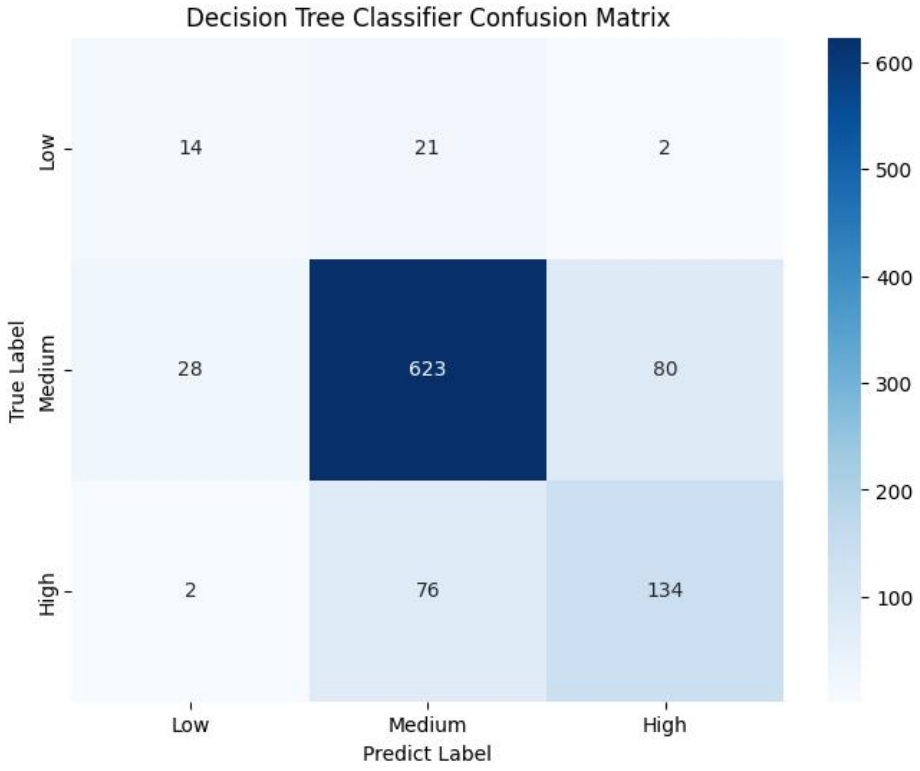
- **Trường hợp train/test = 80/20:**

```
Classification Report:
              precision    recall  f1-score   support

         Low       0.32      0.38      0.35        37
      Medium       0.87      0.85      0.86       731
        High       0.62      0.63      0.63       212

    accuracy                           0.79       980
   macro avg       0.60      0.62      0.61       980
weighted avg       0.79      0.79      0.79       980
```

Trong mô hình có tổng cộng 37 mẫu thuộc lớp Low, 731 mẫu thuộc lớp Medium, 212 mẫu thuộc lớp High, mô hình dự đoán đúng 79% số mẫu trong tập kiểm tra.

- *Lớp Low:*
    - Precision: dự đoán chính xác 32%
    - Recall: Mô hình phát hiện được 38% số mẫu thật sự thuộc lớp Low
    - F1-score: 35% cho thấy hiệu suất cân bằng của lớp Low rất thấp.
- *Lớp Medium:*
    - Precision: dự đoán chính xác 87%
    - Recall: Mô hình phát hiện được 85% số mẫu thật sự thuộc lớp Medium
    - F1-score: 86% cho thấy hiệu suất cân bằng của lớp Medium tốt.
    - Lớp Medium trong mô hình này cũng là lớp đạt được hiệu quả cao nhất với các tỷ lệ trên 80%.
- *Lớp High:*
    - Precision: dự đoán chính xác 62%
    - Recall: Mô hình phát hiện được 63% số mẫu thật sự thuộc lớp High.
    - F1-score: 63% cho thấy hiệu suất cân bằng của lớp High ở mức ổn tuy nhiên đây vẫn chưa phải một kết quả lý tưởng.
- *Các chỉ số tính trung bình:*
    - Macro Average: trung bình giữa ba lớp: precision 0.60, recall 0.62, f1-score 0.61.
    - Weighted Average trung bình nhưng có dựa trên số mẫu của mỗi lớp trong mô hình: precision 0.79, recall 0.79, f1-score 0.79.

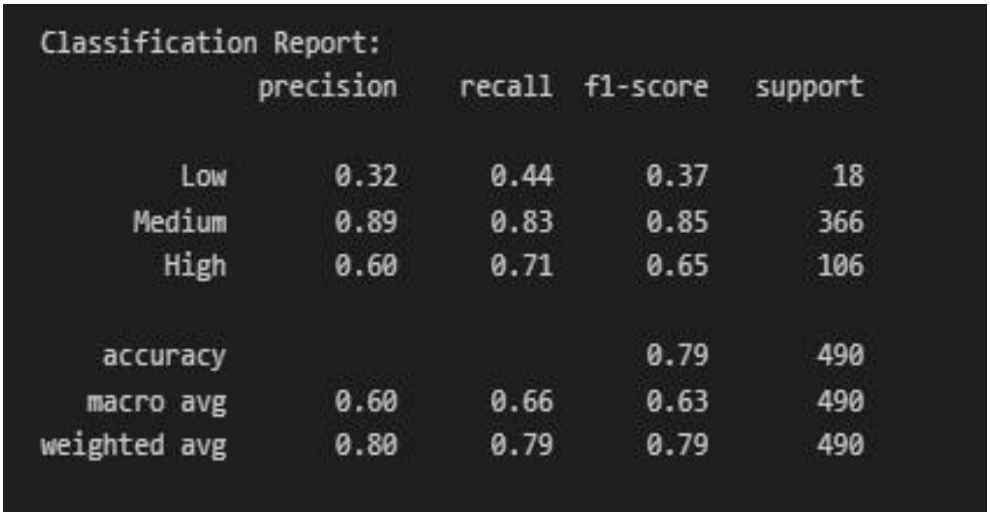
-> Có sự chênh lệch này là do lớp Low đạt độ chính xác thấp nhưng thực tế lớp Low lại có rất ít mẫu.



Nhận xét:

- Mô hình làm việc tốt nhất trong việc phân loại lớp Medium, tuy nhiên mẫu lớp Medium đa số vẫn còn bị nhầm lẫn sang lớp High, có 80 mẫu bị nhầm sang lớp High, và 28 mẫu bị nhầm sang lớp Low.
- Và ngược lại, gần ⅓ mẫu từ lớp High (76 mẫu) bị nhầm lẫn sang lớp Medium.
- Lớp Low có hoạt động dự đoán kém nhất, khi hơn phân nửa số mẫu của lớp Low bị nhầm sang hai lớp còn lại. Tuy nhiên, trong trường hợp với tỷ lệ 80/20 này, lớp Low hoạt động tốt nhất trong các trường hợp.

- **Trường hợp train/test = 90/10:**



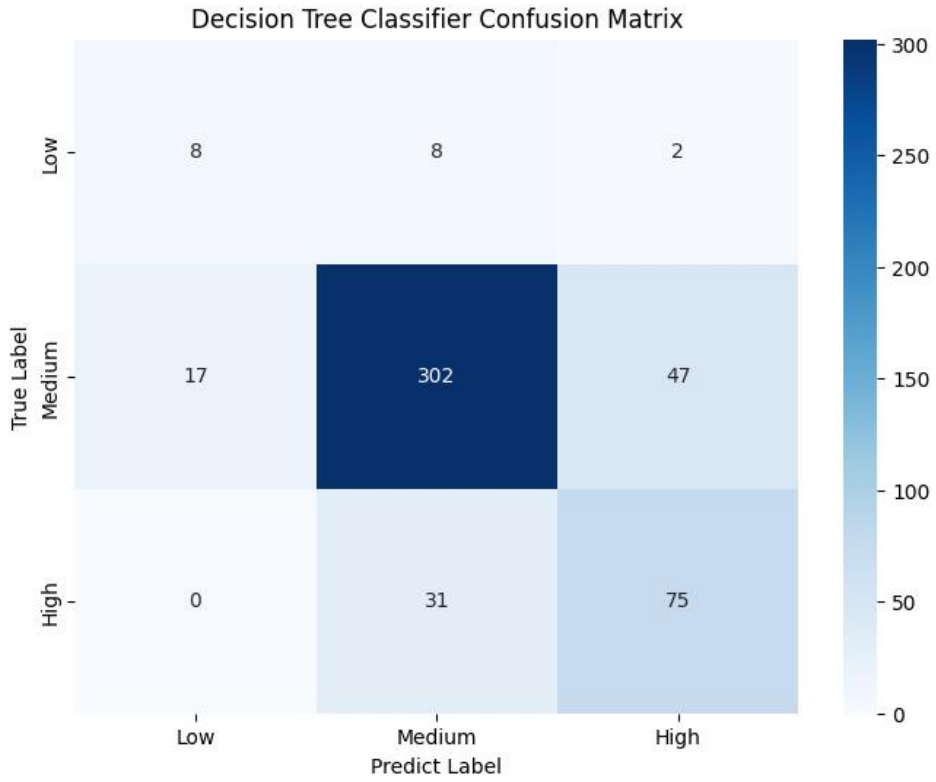
Classification Report:

	precision	recall	f1-score	support
Low	0.32	0.44	0.37	18
Medium	0.89	0.83	0.85	366
High	0.60	0.71	0.65	106
accuracy			0.79	490
macro avg	0.60	0.66	0.63	490
weighted avg	0.80	0.79	0.79	490

Trong mô hình có tổng cộng 18 mẫu thuộc lớp Low, 366 mẫu thuộc lớp Medium, 106 mẫu thuộc lớp High, mô hình dự đoán đúng 79% số mẫu trong tập kiểm tra.

- *Lớp Low:*
  - Precision: dự đoán chính xác 32%
  - Recall: Mô hình phát hiện được 44% số mẫu thật sự thuộc lớp Low
  - F1-score: 37% cho thấy hiệu suất cân bằng của lớp Low rất thấp.
- *Lớp Medium:*
  - Precision: dự đoán chính xác 89%
  - Recall: Mô hình phát hiện được 83% số mẫu thật sự thuộc lớp Medium
  - F1-score: 85% cho thấy hiệu suất cân bằng của lớp Medium tốt.
  - Lớp Medium trong mô hình này cũng là lớp đạt được hiệu quả cao nhất với các tỷ lệ trên 80%.
- *Lớp High:*
  - Precision: dự đoán chính xác 60%
  - Recall: Mô hình phát hiện được 71% số mẫu thật sự thuộc lớp Low
  - F1-score: 65% cho thấy hiệu suất cân bằng của lớp High ở mức ổn tuy nhiên đây vẫn chưa phải một kết quả lý tưởng.
- *Các chỉ số tính trung bình:*
  - Macro Average: trung bình giữa ba lớp: precision 0.6, recall 0.66, f1-score 0.63.
  - Weighted Average trung bình nhưng có dựa trên số mẫu của mỗi lớp trong mô hình: precision 0.8, recall 0.79, f1-score 0.79.

-> Có sự chênh lệch này là do lớp Low đạt độ chính xác thấp nhưng thực tế lớp Low lại có rất ít mẫu.



Nhận xét:

- Mô hình làm việc tốt nhất trong việc phân loại lớp Medium, tuy nhiên mẫu lớp Medium đa số vẫn còn bị nhầm lẫn sang lớp High (47 mẫu bị nhầm sang lớp High, 17 mẫu bị nhầm sang lớp Low).
- Và ngược lại, gần ⅓ mẫu từ lớp High bị nhầm lẫn sang lớp Medium, không nhầm sang lớp Low (31 mẫu).
- Lớp Low có hoạt động dự đoán kém nhất, khi hơn phân nửa số mẫu của lớp Low bị nhầm sang hai lớp còn lại.
- Ta nhận thấy, ma trận nhầm lẫn của trường hợp này tương đối giống trường hợp 80/20, chỉ là kết quả của lớp High được cải tiến.

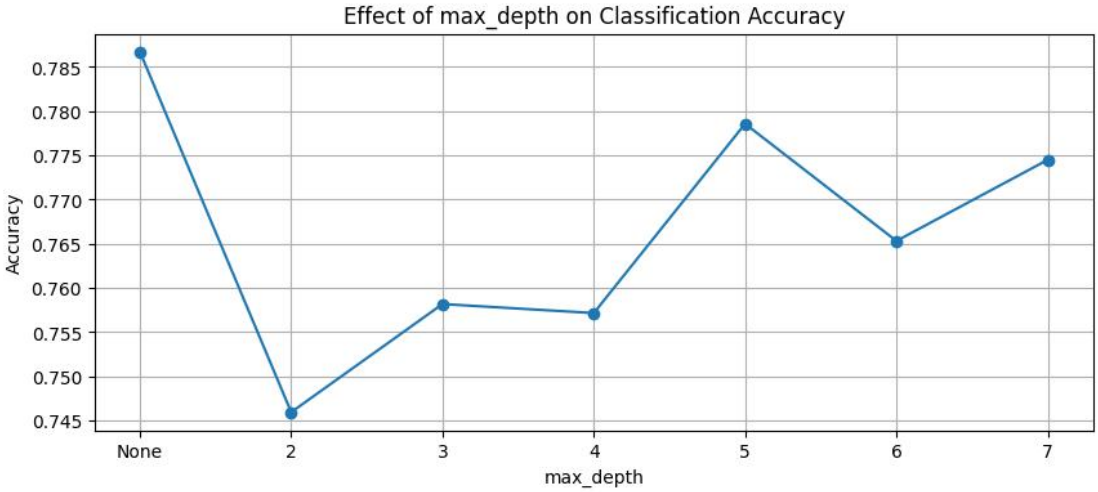
4. Đánh giá độ sâu và độ chính xác của cây quyết định.

- Hình ảnh cây quyết định được lưu trong thư mục **DecisionTree_Images** (thư mục ở cùng địa chỉ với file source code, thư mục xuất hiện sau khi chạy file WineQuality.ipynb).

- Bảng thống kệ độ chính xác của mô hình theo chiều sâu của cây quyết định.

max_depth	None	2	3	4	5	6	7
accuracy	0.786	0.745	0.758	0.757	0.778	0.765	0.774

- Biểu đồ trực quan mối liên hệ giữa độ chính xác của dự đoán và chiều sâu của cây quyết định.



- **Nhận xét:** Nhìn chung, độ chính xác của mô hình qua các cây quyết định có chiều sâu tối đa khác nhau không có khác biệt quá lớn (trong khoảng 74-78%). Đạt độ chính xác cao nhất là cây quyết định với độ sâu không giới hạn với 78.6%, tuy nhiên việc không giới hạn độ sâu tối đa của cây sẽ dẫn đến việc cây rất phức tạp. Đạt độ chính xác thấp nhất là cây có độ sâu tối đa bằng 2 với 74.5%. Ta không thể kết luận độ sâu tối đa của cây quyết định sẽ hoàn toàn quyết định độ chính xác của nó, bởi vì biểu đồ đường bên trên đã cho ta thấy rằng, độ chính xác của mô hình không cùng tăng với độ sâu tối đa của cây quyết định.

b. Breast Cancer dataset

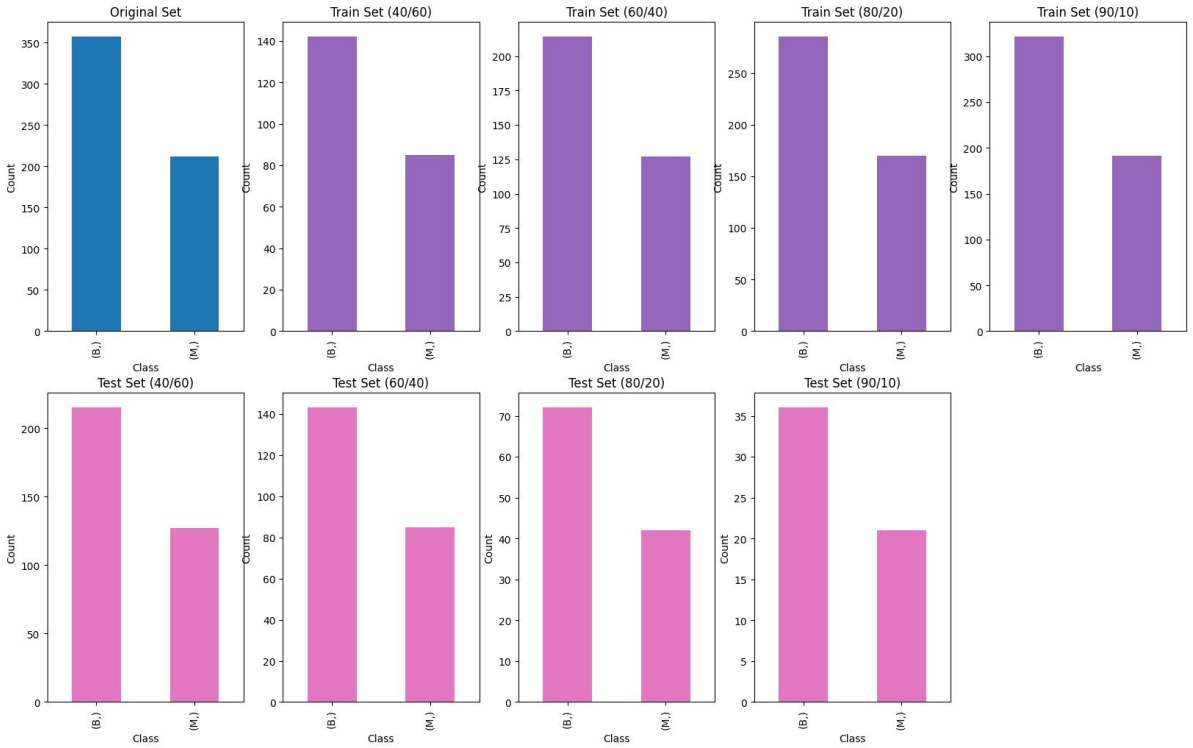
i. Giới thiệu về dữ liệu

- Nguồn: [Breast Cancer Wisconsin (Diagnostic)]
- Bộ dữ liệu bao gồm 569 dòng và 30 cột. Mỗi dòng đại diện cho một khối u, bao gồm các đặc tính hình học mô tả hình dạng và tính chất của khối u.
- Ý nghĩa các cột dữ liệu:
  - **radiusX** (trung bình khoảng cách từ tâm đến các điểm trên viền của nó): bán kính của khối u
  - **textureX**: độ lệch chuẩn của các giá trị độ sáng xám của các pixel trong hình ảnh
  - **perimeterX**: chu vi của khối u
  - **areaX**: diện tích của khối u
  - **smoothnessX** (biến thiên cục bộ trong độ dài bán kính): độ mịn của biên khối u
  - **compactnessX** (chuvi^2 / diện tích - 1.0): độ nén của khối u
  - **concavityX**: mức độ lõm vào của đường biên khối u
  - **concave_pointsX**: số lượng điểm lõm trên đường biên khối u
  - **symmetryX**: độ đối xứng của khối u
  - **fractal_dimensionX** (“xấp xỉ đường bờ biển” - 1): sự phức tạp của hình dạng khối u, liên quan đến các chi tiết nhỏ trên đường biên, sử dụng mô hình toán học được gọi là “xấp xỉ đường bờ biển” để tính toán

30 cột bao gồm 3 nhóm 10 cột đặc trưng như trên nhưng thay X lần lượt là 1, 2, 3, đại diện cho các lần đo, tính toán khác nhau.

ii. Phân tích dữ liệu

1. Chuẩn bị dữ liệu (chia dữ liệu)



- Sau khi chia dữ liệu, tỷ lệ phân bố giữa các lớp vẫn được duy trì đồng đều ở cả tập huấn luyện và kiểm tra. Đảm bảo mô hình được phân tích trên các dữ liệu có đặc điểm tương tự, giúp kết quả phân loại trở nên đáng tin cậy. Việc duy trì sự phân bố hợp lý giữa các lớp trong tất cả các tập dữ liệu chứng minh rằng việc chia dữ liệu đã được thực hiện đúng cách.

2. Xây dựng cây quyết định (Decision Tree)

Hình ảnh cây quyết định được lưu dưới định dạng file .svg trong thư mục Decision_Tree_Visualizations_BreastCancer_Dataset (thư mục có đường dẫn giống với đường dẫn đến file source code).

3. Đánh giá kết quả qua Classification Report và Confusion Matrix.

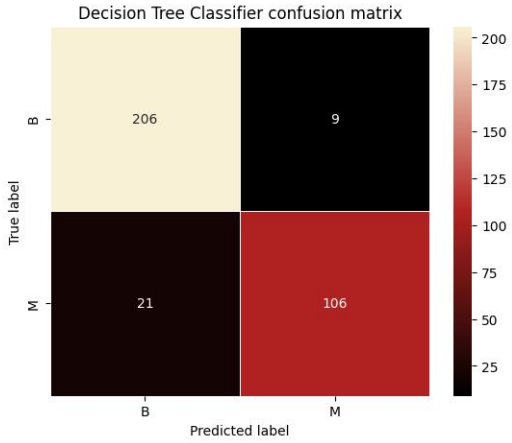
- **Classification Report** hiển thị các chỉ số đánh giá hiệu suất mô hình trên từng lớp của dữ liệu.
    - *precision:* tỷ lệ dự đoán đúng trong số các dự đoán của mỗi lớp.
    - *recall:* tỷ lệ nhận diện đúng mẫu thuộc một lớp nào đó.
    - *f1-score:* kết hợp giữa **precision** và **recall**
    - *support: số lượng mẫu thật sự của mỗi lớp.*
    - *accuracy: độ chính xác của tổng thể mô hình*
    - *macro avg: trung bình của precision, recall và f1-score*
    - *weighted avg: trung bình dựa trên số lượng mẫu trong mỗi lớp*

- **Confusion Matrix** là ma trận nhầm lẫn thể hiện các lỗi trong dự đoán của mô hình.
    - Các phần tử nằm trên *đường chéo chính* của ma trận sẽ biểu thị số lượng các dự đoán chính xác.

    - Các phần tử còn lại là số dự đoán sai, bị nhầm từ lớp i sang lớp j. (với j là chỉ số cột, i là chỉ số dòng)

- **Trường hợp train/test = 40/60:**

	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



Kết quả từ mô hình cho thấy hiệu suất khá tốt trong việc phân loại các khối u thành 2 lớp. Dữ liệu bao gồm 215 mẫu thuộc lớp Benign (lành tính) và 127 mẫu thuộc lớp Malignant (ác tính).

- Lớp Benign (B): mô hình đạt precision 0.91 và recall 0.96 cho thấy mô hình có khả năng phân loại chính xác các khối u lành tính, hầu như không bỏ sót mẫu nào (vô cùng ít). f1-score 0.93 thể hiện sự cân bằng tốt giữa precision và recall, với việc mô hình nhận diện tốt các mẫu Benign mà không bị nhầm lẫn nhiều.

- Lớp Malignant (M): mô hình đạt precision 0.92 và recall 0.83 mặc dù có tỷ lệ chính xác cao trong việc loại các khối u khác tính, nhưng vẫn có khoảng 17% mẫu Malignant bị bỏ sót, dẫn đến recall thấp hơn so với lớp Benign, f1-score 0.88 cho thấy mô hình vẫn có hiệu suất khá tốt.

- Mô hình đạt Accuracy 91% nghĩa là có 91% tổng số mẫu được phân loại chính xác cho thấy mô hình học được thông tin cần thiết từ dữ liệu

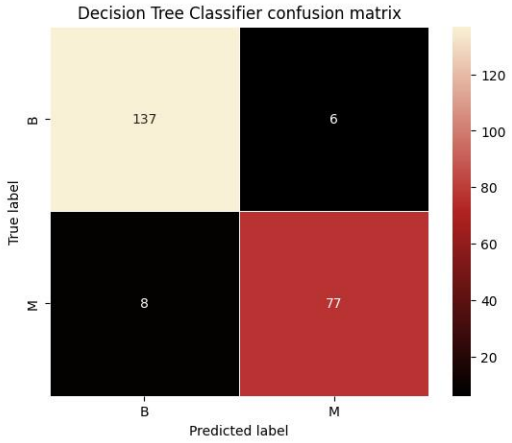
- Macro avg: Precision và Recall trung bình trên tất cả các lớp đều ở mức 0.91 và 0.90, cho thấy mô hình phân loại khá đồng đều giữa hai lớp.

- Weighted avg: Giá trị F1-score và Accuracy có trọng số trung bình là 0.91, cũng cho thấy hiệu suất tổng thể ổn định.

- Confusion matrix: cho thấy mô hình dự đoán đúng 206 mẫu Benign và 106 mẫu Malignant. Tuy nhiên, vẫn có một số mẫu bị phân loại sai: 9 mẫu Benign bị nhầm là Malignant và 21 mẫu Malignant bị nhầm là Benign. Điều này phản ánh rằng mô hình có thể cần cải thiện trong việc nhận diện đầy đủ các mẫu Malignant.

- **Trường hợp train/test = 60/40:**

	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



Kết quả từ mô hình phân loại cho thấy hiệu suất của mô hình khá tốt khi phân loại các khối u thành hai lớp Benign (lành tính) và Malignant (ác tính), với dữ liệu bao gồm 143 mẫu lớp Benign và 85 mẫu lớp Malignant.

- Lớp Benign (B): Mô hình đạt precision 0.94 và recall 0.96, cho thấy khả năng phân loại chính xác các khối u lành tính rất cao, mô hình hầu như không bỏ sót mẫu nào, f1-score 0.95 cho thấy sự cân bằng rất tốt giữa precision và recall, thể hiện hiệu quả trong việc nhận diện các mẫu Benign.

- Lớp Malignant (M): Mô hình đạt precision 0.93 và recall 0.91 cho thấy dù mô hình có precision cao, có nghĩa là nó phân loại chính xác được các mẫu Malignant, tuy nhiên recall chỉ đạt 91%, cho thấy vẫn có một số mẫu Malignant bị bỏ sót, f1-score là 0.92, cho thấy mô hình có hiệu suất khá ổn định trong việc phân loại các khối u ác tính.

- Accuracy: Mô hình đạt Accuracy 94%, nghĩa là có 94% tổng số mẫu được phân loại chính xác.

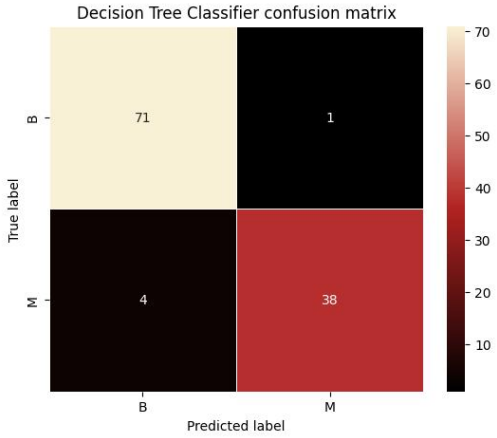
- Macro avg: precision và recall trung bình trên tất cả các lớp lần lượt là 0.94 và 0.93, cho thấy mô hình phân loại khá đồng đều giữa hai lớp.

- Weighted avg: Giá trị f1-score và Accuracy có trọng số trung bình là 0.94, cho thấy hiệu suất tổng thể của mô hình khá ổn định và mạnh mẽ.

- Confusion Matrix: Mô hình dự đoán đúng 137 mẫu Benign và 77 mẫu Malignant. Tuy nhiên, vẫn có một số mẫu bị phân loại sai: 6 mẫu Benign bị nhầm là Malignant và 8 mẫu Malignant bị nhầm là Benign. Mặc dù tỷ lệ nhầm là khá thấp, nhưng việc giảm thiểu sai sót này sẽ giúp cải thiện hiệu quả nhận diện các khối u ác tính.

- **Trường hợp train/test = 80/20:**

	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



Kết quả từ mô hình phân loại cho thấy hiệu suất của mô hình rất tốt trong việc phân loại các khối u thành hai lớp Benign (lành tính) và Malignant (ác tính), với dữ liệu bao gồm 72 mẫu lớp Benign và 42 mẫu lớp Malignant.

- Lớp Benign (B): Mô hình đạt precision 0.95 và recall 0.99, cho thấy mô hình có khả năng phân loại chính xác các khối u lành tính rất cao và hầu như không bỏ sót mẫu nào, recall cực kỳ cao lên đến 99%, f1-score là 0.97, thể hiện sự cân bằng rất tốt giữa precision và recall, mô hình rất hiệu quả trong việc phân loại các mẫu Benign.

- Lớp Malignant (M): Mô hình đạt precision 0.97 và recall 0.90. Mặc dù mô hình có precision cao, tức là khả năng phân loại chính xác các khối u ác tính là rất tốt, f1-score là 0.94, thể hiện hiệu suất khá tốt trong việc phân loại các khối u ác tính.

- Accuracy: Mô hình đạt Accuracy 96%, nghĩa là 96% tổng số mẫu được phân loại chính xác, thể hiện mô hình phân loại rất hiệu quả.

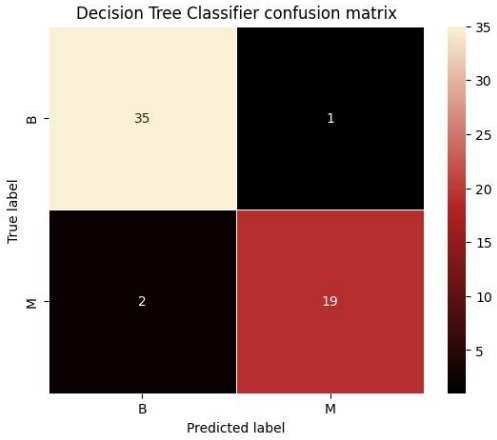
- Macro avg: precision và recall trung bình trên tất cả các lớp lần lượt là 0.96 và 0.95, cho thấy mô hình phân loại khá đồng đều giữa hai lớp.

- Weighted avg: Giá trị f1-score và Accuracy có trọng số trung bình là 0.96, cho thấy hiệu suất tổng thể của mô hình rất ổn định và mạnh mẽ.

- Confusion Matrix: Mô hình dự đoán đúng 71 mẫu Benign và 38 mẫu Malignant. Tuy nhiên vẫn có một số mẫu bị phân loại sai: 1 mẫu Benign bị nhầm là Malignant và 4 mẫu Malignant bị nhầm là Benign (False Negative) nhưng mà tỷ lệ sai sót này là đã cực kỳ thấp.

- **Trường hợp train/test = 90/10:**

	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



Kết quả từ mô hình phân loại cho thấy hiệu suất của mô hình khá tốt trong việc phân loại các khối u thành hai lớp Benign (lành tính) và Malignant (ác tính), với dữ liệu bao gồm 36 mẫu lớp Benign và 21 mẫu lớp Malignant.

- Lớp Benign (B): Mô hình đạt precision 0.95 và recall 0.97, cho thấy mô hình có khả năng phân loại chính xác các khối u lành tính rất cao, f1-score là 0.96, thể hiện sự cân bằng tốt giữa precision và recall, mô hình hiệu quả trong việc phân loại các mẫu Benign.

- Lớp Malignant (M): Mô hình đạt precision 0.95 và recall 0.90. Precision cao cho thấy mô hình phân loại chính xác các khối u ác tính rất tốt, f1-score là 0.93, thể hiện hiệu suất khá tốt trong việc phân loại khối u ác tính.

- Accuracy: Mô hình đạt Accuracy 95%, tức là 95% tổng số mẫu được phân loại chính xác, cho thấy mô hình phân loại rất hiệu quả.

- Macro avg: precision và recall trung bình trên tất cả các lớp lần lượt là 0.95 và 0.94, cho thấy mô hình phân loại khá đồng đều giữa hai lớp.

- Weighted avg: Giá trị f1-score và Accuracy có trọng số trung bình là 0.95, cho thấy hiệu suất tổng thể ổn định và mạnh mẽ của mô hình.

- Confusion Matrix: Mô hình dự đoán đúng 35 mẫu Benign và 19 mẫu Malignant. Tuy nhiên vẫn có một số mẫu bị phân loại sai: 1 mẫu Benign bị nhầm là Malignant và 2 mẫu Malignant bị nhầm là Benign, nhưng tỷ lệ sai sót đã là cực kỳ thấp.

=> Nhìn chung thì trường hợp train/test = 80/20 là tốt nhất, đạt độ chính xác cao nhất, 90/10 cũng tốt nhưng mà recall của lớp Malignant thì thấp hơn một chút, 60/40 thì ổn định, chênh lệch giữa precision và recall ở lớp Malignant rõ hơn, trường hợp train/test = 40/60 có hiệu suất thấp nhất do recall của lớp Malignant thấp hơn các trường hợp khác đáng kể, dễ bỏ sót các mẫu ác tính.

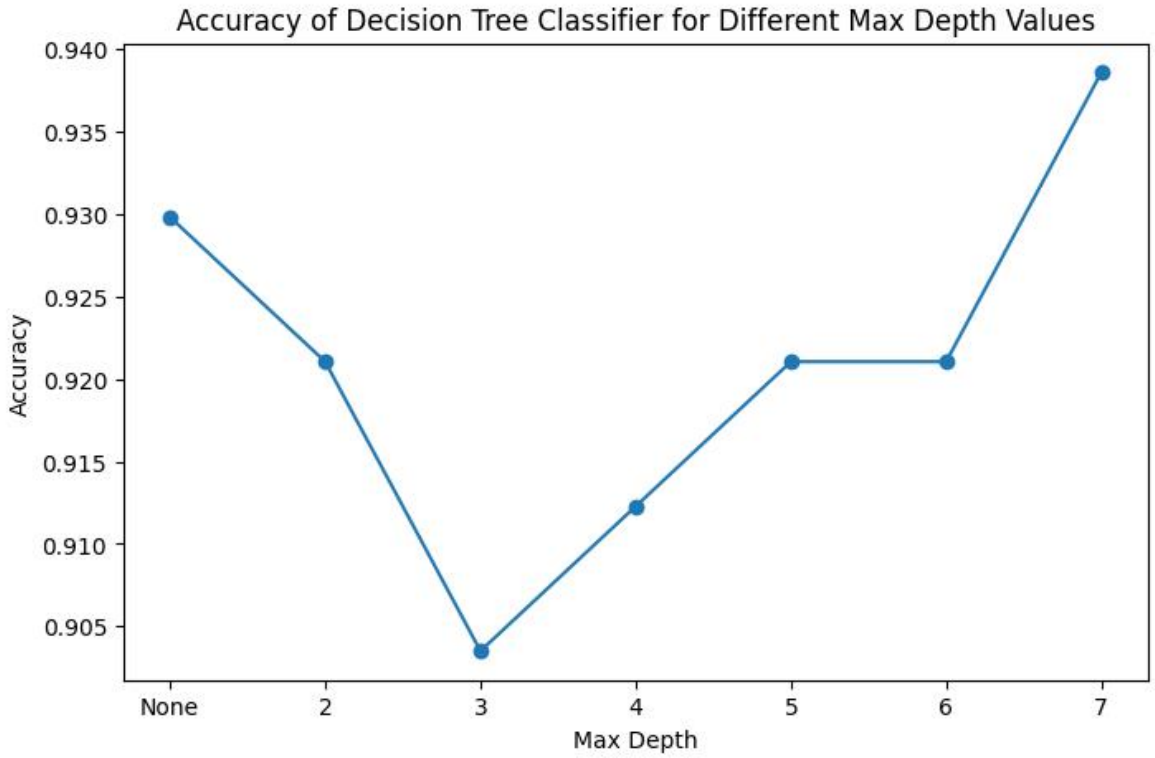
4. Đánh giá độ sâu và độ chính xác của cây quyết định.

- Hình ảnh cây quyết định được lưu trong thư mục Breast_Cancer thuộc thư mục DecisionTree_Images (thư mục ở cùng địa chỉ với file source code).

- Bảng thống kê độ chính xác của mô hình theo chiều sâu của cây quyết định.

max_depth	None	2	3	4	5	6	7
Accuracy	0.9298	0.9211	0.9035	0.9123	0.9211	0.9211	0.9386

- Biểu đồ trực quan mối liên hệ giữa độ chính xác của dự đoán và chiều sâu của cây quyết định.



- **Nhận xét**: khi không giới hạn về độ sâu, cây có khả năng học tối đa từ dữ liệu, Accuracy cao (0.9298) nhưng lại không cao bằng cây khi có độ sâu bằng 7 (0.9386), điều này cho thấy khi giới hạn độ sâu ở mức hợp lý, cây quyết định đạt độ chính xác tốt hơn, là lựa chọn cân bằng giữa hiệu quả và độ phức tạp, vừa học được thông tin từ dữ liệu vừa tránh overfitting. Giá trị Accuracy thấp nhất tại độ sâu là 3 (0.9035), bởi vì khi độ sâu tối đa bị giới hạn quá nhỏ thì cây không đủ phức tạp để học mối quan hệ trong dữ liệu. Sở dĩ ở độ sâu bằng 3 nó không tốt bằng 2 là vì bài toán không quá phức tạp nên ở độ sâu bằng 2, các đặc trưng quan trọng đã đủ để phân loại phần lớn dữ liệu, cây đơn giản không bị ảnh hưởng bởi nhiễu hoặc các thứ khác. Độ sâu bằng 3 sẽ bị ảnh hưởng nhiều hơn và nó cũng không tốt bằng 4, 5, 6 vì ở những độ sâu này có thể khai thác thêm được các đặc trưng quan trọng khác.

c. Additional dataset

i. Giới thiệu về dữ liệu

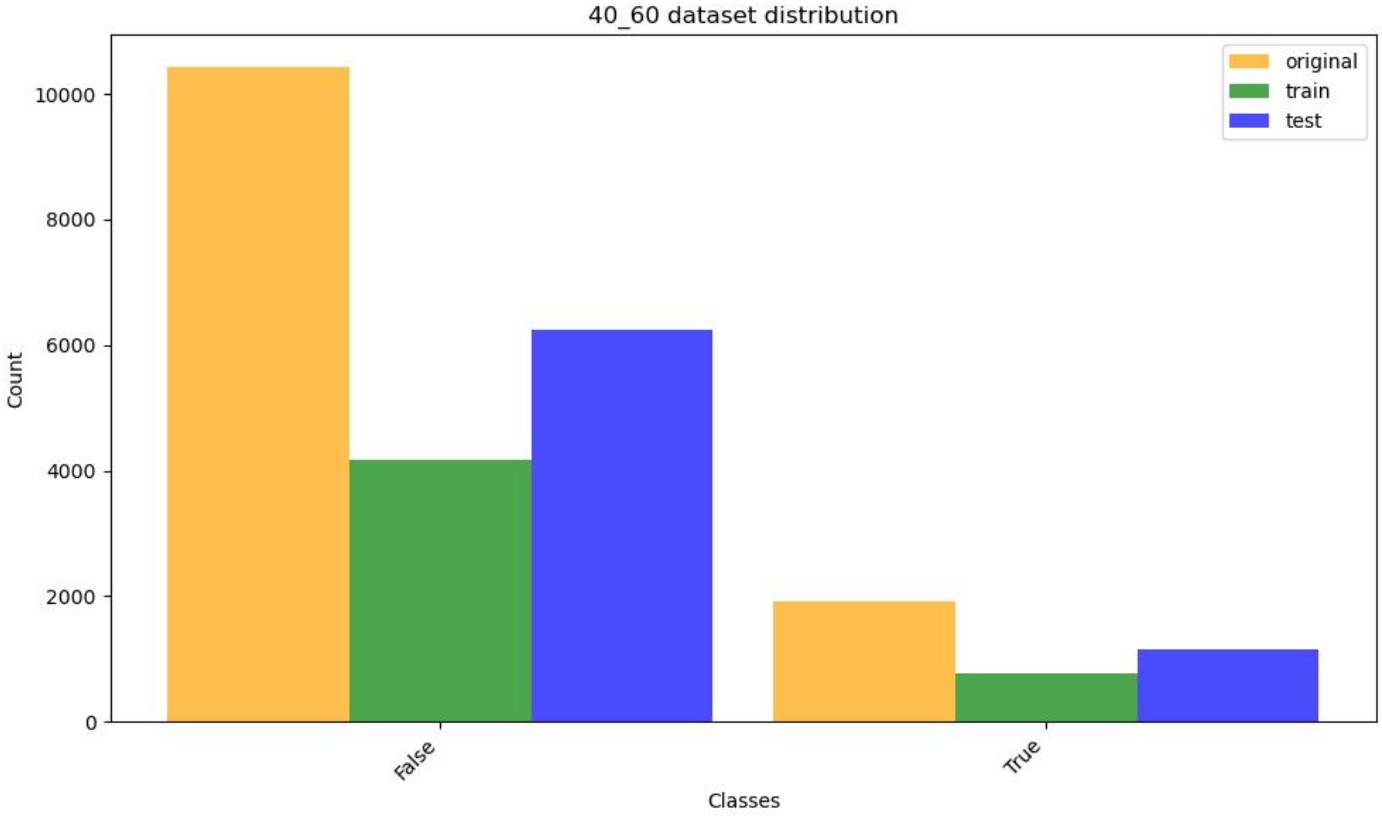
- Nguồn: [Online Shoppers Purchasing Intention Dataset]

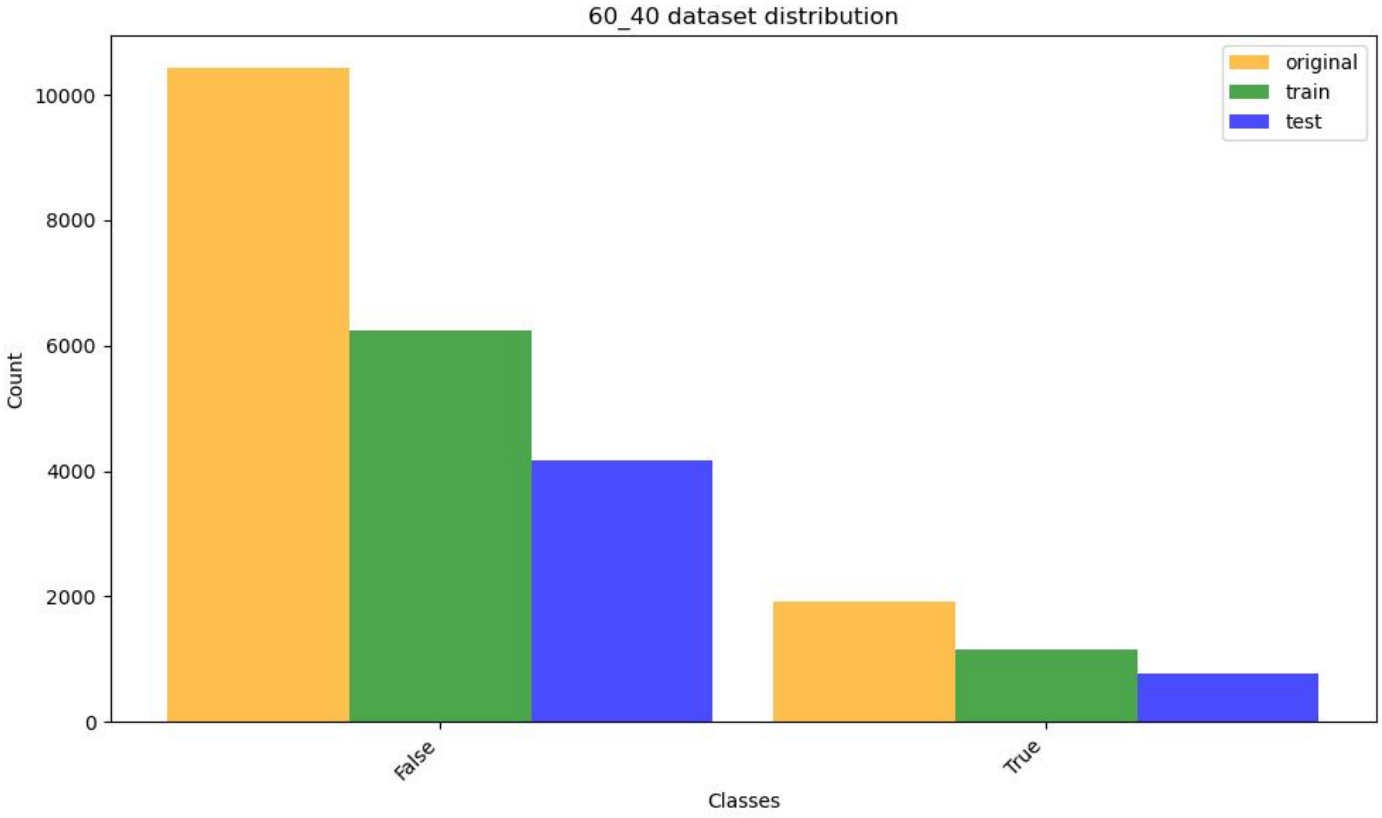
- Bộ dữ liệu bao gồm các vector đặc trưng thuộc về 12.330 sessions. Bộ dữ liệu được xây dựng sao cho mỗi session thuộc về một người dùng khác nhau trong khoảng thời gian 1 năm, nhằm tránh xu hướng thiên lệch theo một chiến dịch cụ thể, ngày đặc biệt, hồ sơ người dùng, hoặc giai đoạn thời gian nào đó.

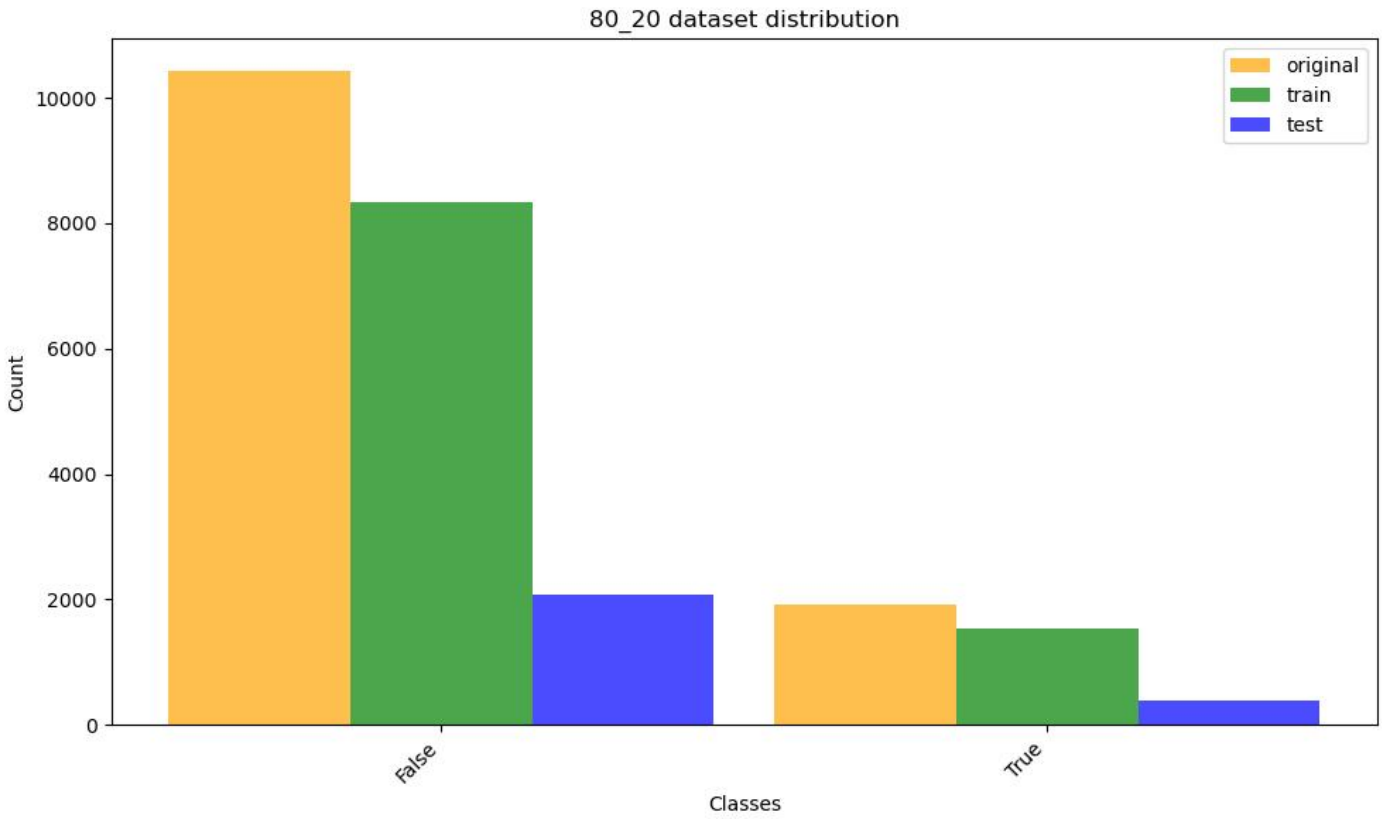
- Ý nghĩa từng cột:
  - **Administrative**: Số lượng các trang thuộc loại hành chính mà người dùng đã truy cập trên trang web.
  - **Administrative_Duration**: ổng thời gian mà người dùng dành trên các trang hành chính.
  - **Informational**: Số lượng các trang cung cấp thông tin mà người dùng đã truy cập.
  - **Informational_Duration**: Tổng thời gian người dùng dành trên các trang thông tin.
  - **ProductRelated**: Số lượng các trang liên quan đến sản phẩm mà người dùng đã duyệt.
  - **ProductRelated_Duration**: Tổng thời gian người dùng dành trên các trang liên quan đến sản phẩm.
  - **BounceRates**: Tỷ lệ thoát trang, tức là tỷ lệ người dùng rời khỏi trang web mà không thực hiện thêm bất kỳ hành động nào.
  - **ExitRates**: Tỷ lệ người dùng rời khỏi một trang cụ thể trên tổng số lượt truy cập trang đó.
  - **PageValues**: Giá trị trung bình của một trang khi xem xét lợi ích doanh thu mà trang đó tạo ra.
  - **SpecialDay**: Điểm đánh giá mức độ gần với các ngày đặc biệt (ví dụ: Black Friday, Lễ Tết).
  - **Month**: Tháng mà hành vi truy cập được ghi nhận.
  - **OperatingSystems**: Hệ điều hành của thiết bị mà người dùng sử dụng.
  - **Browser**: Trình duyệt mà người dùng sử dụng để truy cập trang web.
  - **Region**: Khu vực địa lý mà người dùng truy cập từ đó.
  - **TrafficType**: Loại lưu lượng truy cập dẫn người dùng đến trang web.
  - **VisitorType**: Loại khách truy cập (Returning_Visitor hoặc New_Visitor).
  - **Weekend**: Người dùng truy cập trang web vào cuối tuần hay không.
  - **Revenue**: Người dùng có tạo ra doanh thu hay không.

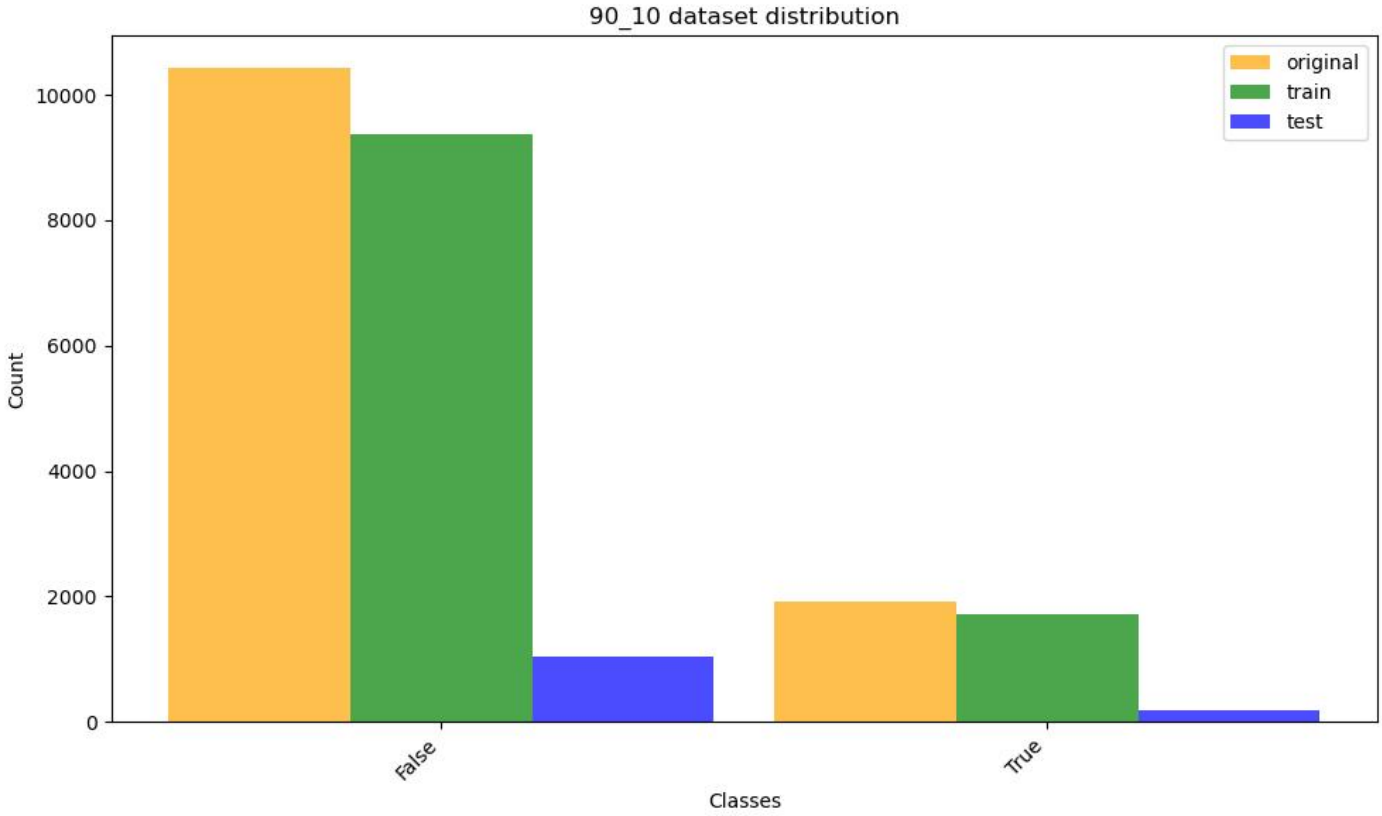
ii. Phân tích dữ liệu

1. Chuẩn bị dữ liệu:









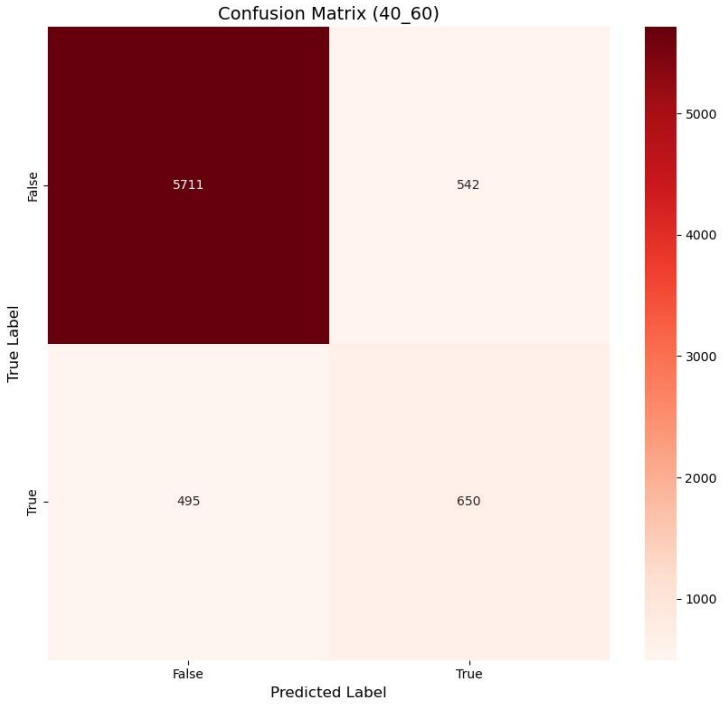
Việc chia tỷ lệ **train/test** giúp đảm bảo mô hình học đủ thông tin từ dữ liệu, đồng thời cung cấp đủ cơ sở để đánh giá hiệu suất trên dữ liệu mới, qua đó tránh được hiện tượng **overfitting** hoặc **underfitting**.

2. Xây dựng cây quyết định

Cây quyết định được tạo ra thành file svg nằm trong thư mục Decision_Tree_Visualizations_Additional_Dataset được tạo ra sau khi chạy hàm train_and_visualize_decision_tree. Hình ảnh được tạo ra nó sẽ tạo ra với format là decision_tree_{split_key} với split key là chia train test dataset thành 40/60, 60/40, 80/20 và 90/10.

3. **Đánh giá kết quả qua Classification report và Confusion Matrix**.
   - **Classification Report** hiển thị các chỉ số đánh giá hiệu suất mô hình trên từng lớp của dữ liệu.
     - *precision:* tỷ lệ dự đoán đúng trong số các dự đoán của mỗi lớp.
     - *recall:* tỷ lệ nhận diện đúng mẫu thuộc một lớp nào đó.
     - *f1-score:* kết hợp giữa **precision** và **recall**
     - *support: số lượng mẫu thật sự của mỗi lớp.*
     - *accuracy: độ chính xác của tổng thể mô hình*
     - *macro avg: trung bình của precision, recall và f1-score*
     - *weighted avg: trung bình dựa trên số lượng mẫu trong mỗi lớp*
   - **Confusion Matrix** là ma trận nhầm lẫn thể hiện các lỗi trong dự đoán của mô hình.
     - Các phần tử nằm trên *đường chéo chính* của ma trận sẽ biểu thị số lượng các dự đoán chính xác.
     - Các phần tử còn lại là số dự đoán sai, bị nhầm từ lớp i sang lớp j. (với j là chỉ số cột, i là chỉ số dòng)

- Tập train test dataset 40/60

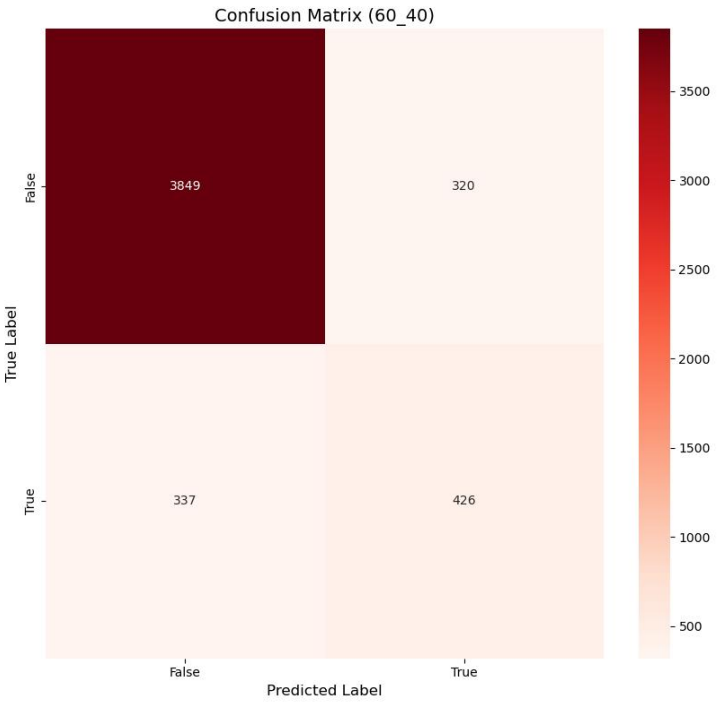


Classification Report (40_60)

	precision	recall	f1-score	support
False	0.92	0.91	0.92	6253
True	0.55	0.57	0.56	1145
accuracy			0.86	7398
macro avg	0.73	0.74	0.74	7398
weighted avg	0.86	0.86	0.86	7398

- **Nhận xét về Confusion Matrix:**
    1. **Phân loại "False"** (True Label = False):
        - Dự đoán đúng: 5711 trường hợp (ô trên bên trái).
        - Dự đoán sai: 542 trường hợp bị nhầm thành "True" (ô trên bên phải).
        - Điều này cho thấy mô hình phân loại khá tốt đối với nhãn "False", nhưng vẫn có một lượng nhỏ nhầm lẫn.
    2. **Phân loại "True"** (True Label = True):
        - Dự đoán đúng: 650 trường hợp (ô dưới bên phải).
        - Dự đoán sai: 495 trường hợp bị nhầm thành "False" (ô dưới bên trái).
        - Kết quả cho thấy mô hình khó phân biệt hơn đối với nhãn "True", dẫn đến số lượng dự đoán sai tương đối cao.
- **Nhận xét về Classification Report:**
    1. **Precision (Độ chính xác)**:
        - Nhãn "False": 0.92 - Rất cao, hầu hết các dự đoán "False" là chính xác.
        - Nhãn "True": 0.55 - Khá thấp, cho thấy mô hình có nhiều dự đoán nhầm với nhãn này.
    2. **Recall (Độ nhạy)**:
        - Nhãn "False": 0.91 - Mô hình nhận diện gần hết các trường hợp "False".
        - Nhãn "True": 0.57 - Khá thấp, mô hình bỏ lỡ nhiều trường hợp thuộc nhãn "True".
    3. **F1-Score**:
        - Nhãn "False": 0.92 - Tốt, cho thấy sự cân bằng giữa Precision và Recall.
        - Nhãn "True": 0.56 - Yếu, mô hình cần cải thiện hiệu suất cho nhãn này.
    4. **Accuracy (Độ chính xác tổng quan)**:
        - Đạt 0.86 (86%), cho thấy mô hình hoạt động khá tốt, nhưng sự chênh lệch hiệu suất giữa hai nhãn là đáng kể.
    5. **Macro Average**:
        - **Precision/Recall/F1**: 0.73/0.74/0.74 → Trung bình hiệu suất giữa hai nhãn có sự chênh lệch lớn do nhãn "True" có kết quả kém hơn.
    6. **Weighted Average**:
        - **Precision/Recall/F1**: 0.86 → Trung bình trọng số tính theo số lượng mẫu, chủ yếu bị chi phối bởi nhãn "False".

- Tập train test dataset 60/40

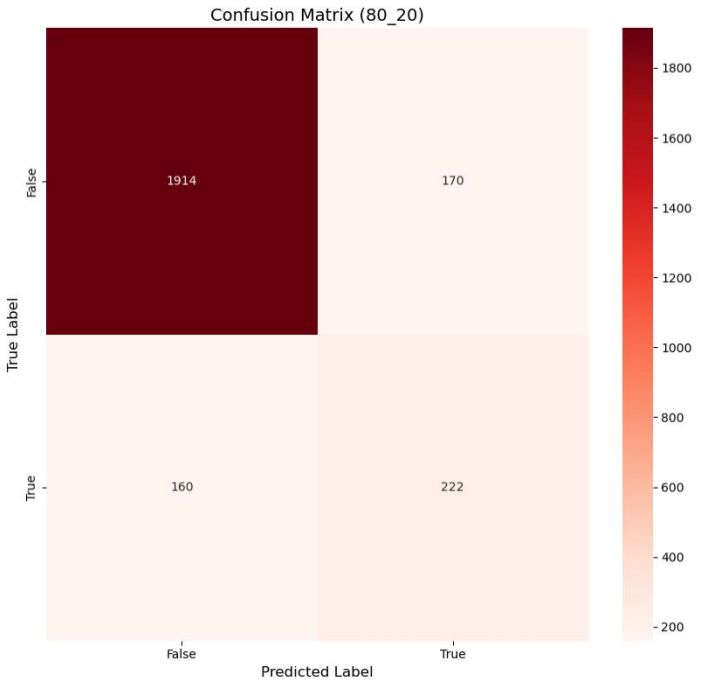


Classification Report (60_40)

	precision	recall	f1-score	support
False	0.92	0.92	0.92	4169
True	0.57	0.56	0.56	763
accuracy			0.87	4932
macro avg	0.75	0.74	0.74	4932
weighted avg	0.87	0.87	0.87	4932

- **Nhận xét Confusion Matrix (60_40)**
  1. **Phân loại "False"** (True Label = False):
     - **Dự đoán đúng**: 3849 trường hợp (ô trên bên trái).
     - **Dự đoán sai**: 320 trường hợp bị nhầm thành "True" (ô trên bên phải).
  2. **Phân loại "True"** (True Label = True):
     - **Dự đoán đúng**: 426 trường hợp (ô dưới bên phải).
     - **Dự đoán sai**: 337 trường hợp bị nhầm thành "False" (ô dưới bên trái).
- **Nhận xét Classification Report (60_40)**
  1. **Precision (Độ chính xác):**
     - **Nhãn "False"**: 0.92 - Rất cao, hầu hết các dự đoán "False" là chính xác.
     - **Nhãn "True"**: 0.57 - Khá thấp, cho thấy mô hình có nhiều dự đoán nhầm đối với nhãn này.
  2. **Recall (Độ nhạy):**
     - **Nhãn "False"**: 0.92 - Mô hình nhận diện gần như đầy đủ các trường hợp "False".
     - **Nhãn "True"**: 0.56 - Khá thấp, mô hình bỏ lỡ nhiều trường hợp thuộc nhãn "True".
  3. **F1-Score:**
     - **Nhãn "False"**: 0.92 - Tốt, thể hiện sự cân bằng giữa Precision và Recall.
     - **Nhãn "True"**: 0.56 - Kém, cần cải thiện hiệu suất cho nhãn này.
  4. **Accuracy (Độ chính xác tổng quan):**
     - Đạt **0.87 (87%)**, cho thấy mô hình hoạt động khá tốt. Tuy nhiên, có sự **chênh lệch đáng kể** về hiệu suất giữa nhãn "False" và "True".
  5. **Macro Average**:
     - **Precision/Recall/F1**: 0.75/0.74/0.74 → Trung bình hiệu suất giữa hai nhãn có sự chênh lệch lớn do nhãn "True" có kết quả kém hơn.
  6. **Weighted Average**:
     - **Precision/Recall/F1**: 0.87 → Trung bình trọng số tính theo số lượng mẫu, chủ yếu bị chi phối bởi nhãn "False".

- Tập train test dataset 80/20



Classification Report (80_20)

	precision	recall	f1-score	support
False	0.92	0.92	0.92	2084
True	0.57	0.58	0.57	382
accuracy			0.87	2466
macro avg	0.74	0.75	0.75	2466
weighted avg	0.87	0.87	0.87	2466

Nhận xét về Confusion Matrix:

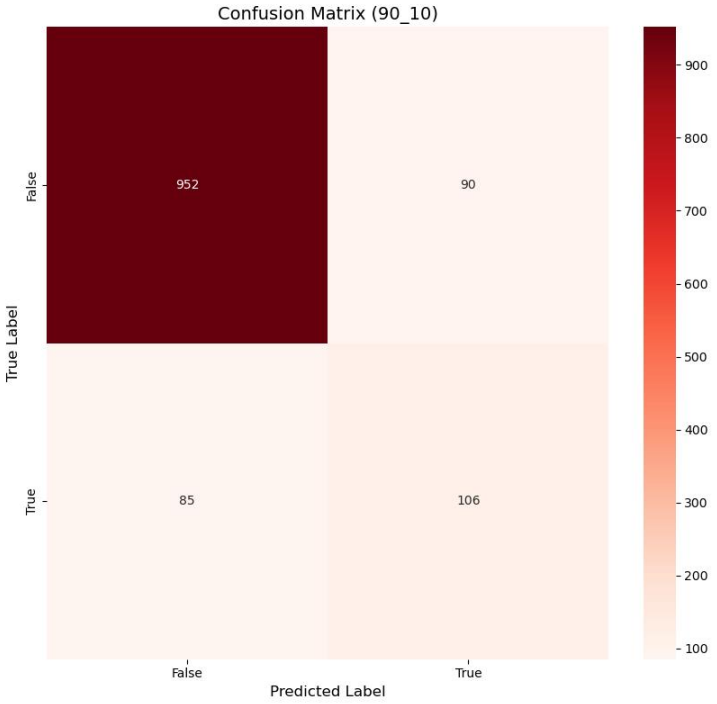
1. **Phân loại "False"**:
   - **Dự đoán đúng**: 1914 trường hợp (ô trên bên trái).
   - **Dự đoán sai**: 170 trường hợp bị nhầm thành "True" (ô trên bên phải).
2. **Phân loại "True"**:
   - **Dự đoán đúng**: 222 trường hợp (ô dưới bên phải).
   - **Dự đoán sai**: 160 trường hợp bị nhầm thành "False" (ô dưới bên trái).

Nhận xét về Classification Report:

1. **Precision (Độ chính xác)**:
   - **Nhãn "False"**: 0.92 - Mô hình có độ chính xác cao khi dự đoán nhãn "False".
   - **Nhãn "True"**: 0.57 - Độ chính xác khi dự đoán nhãn "True" còn thấp, có nhiều trường hợp bị nhầm lẫn.
2. **Recall (Độ nhạy)**:
   - **Nhãn "False"**: 0.92 - Mô hình nhận diện phần lớn các trường hợp "False" một cách chính xác.
   - **Nhãn "True"**: 0.58 - Mô hình bỏ lỡ một phần đáng kể các trường hợp "True".
3. **F1-Score**:
   - **Nhãn "False"**: 0.92 - Hiệu suất của mô hình trên nhãn này rất tốt.
   - **Nhãn "True"**: 0.57 - Mức độ cân bằng giữa Precision và Recall của nhãn này còn khá kém.
4. **Độ chính xác tổng thể (Accuracy)**:
   - **Đạt 87%** - Mô hình có hiệu suất khá tốt khi xét trên toàn bộ tập dữ liệu.
5. **Macro Average**:
   - **Precision/Recall/F1**: 0.74/0.75/0.75 → Trung bình hiệu suất giữa hai nhãn có sự chênh lệch lớn do nhãn "True" có kết quả kém hơn.

6. **Weighted Average**:
    - **Precision/Recall/F1**: 0.87 → Trung bình trọng số tính theo số lượng mẫu, chủ yếu bị chi phối bởi nhãn "False".

- Tập train test dataset 90/10



Classification Report (90_10)

	precision	recall	f1-score	support
False	0.92	0.91	0.92	1042
True	0.54	0.55	0.55	191
accuracy			0.86	1233
macro avg	0.73	0.73	0.73	1233
weighted avg	0.86	0.86	0.86	1233

Nhận xét về Confusion Matrix:

1. **Phân loại "False"**:
    - **Dự đoán đúng**: 952 trường hợp (ô trên bên trái).
    - **Dự đoán sai**: 90 trường hợp bị nhầm thành "True" (ô trên bên phải).
2. **Phân loại "True"**:
    - **Dự đoán đúng**: 106 trường hợp (ô dưới bên phải).
    - **Dự đoán sai**: 85 trường hợp bị nhầm thành "False" (ô dưới bên trái).

Nhận xét về Classification Report:

1. **Precision (Độ chính xác)**:
    - **Nhãn "False"**: 0.92 - Mô hình có độ chính xác cao khi dự đoán nhãn "False".
    - **Nhãn "True"**: 0.54 - Độ chính xác khi dự đoán nhãn "True" còn thấp, có nhiều trường hợp bị nhầm lẫn.
2. **Recall (Độ nhạy)**:
    - **Nhãn "False"**: 0.91 - Mô hình nhận diện phần lớn các trường hợp "False" một cách chính xác.
    - **Nhãn "True"**: 0.55 - Mô hình bỏ lỡ một phần đáng kể các trường hợp "True".
3. **F1-Score**:
    - **Nhãn "False"**: 0.92 - Hiệu suất của mô hình trên nhãn này rất tốt.
    - **Nhãn "True"**: 0.55 - Mức độ cân bằng giữa Precision và Recall của nhãn này còn khá kém.
4. **Độ chính xác tổng thể (Accuracy)**:
    - **Đạt 86%** - Mô hình có hiệu suất khá tốt khi xét trên toàn bộ tập dữ liệu.

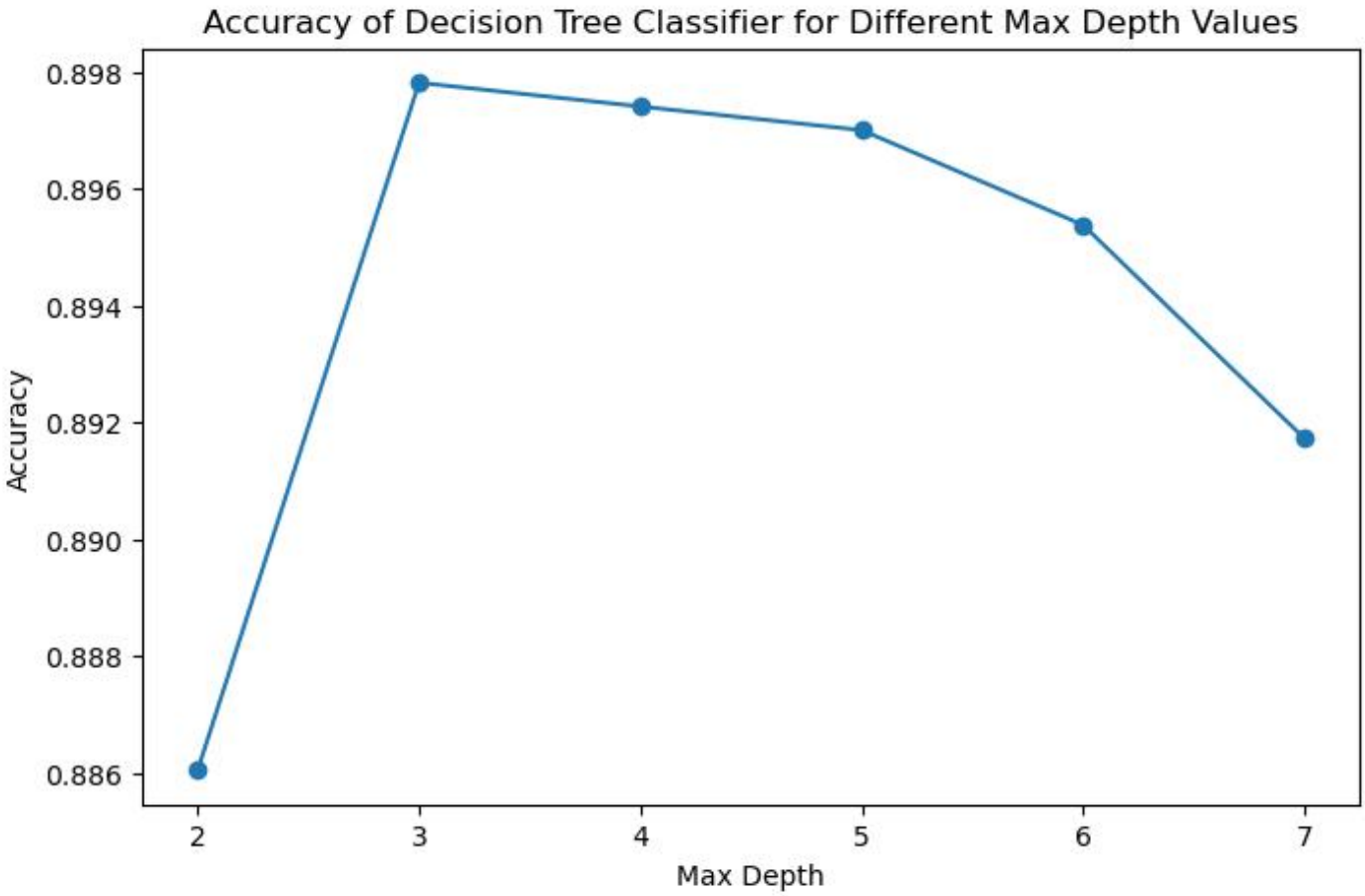
5. **Macro Average**:
    - **Precision/Recall/F1**: 0.73/0.73/0.73 → Trung bình hiệu suất giữa hai nhãn có sự chênh lệch lớn do nhãn "True" có kết quả kém hơn.
6. **Weighted Average**:
    - **Precision/Recall/F1**: 0.86 → Trung bình trọng số tính theo số lượng mẫu, chủ yếu bị chi phối bởi nhãn "False".

4. Đánh giá độ sâu và độ chính xác của cây quyết định.

- Cây quyết định được tạo ra thành file svg nằm trong thư mục Decision_Tree_Visualizations_Additional_Dataset được tạo ra sau khi chạy hàm train_visualize_depth. Hình ảnh được tạo ra nó sẽ tạo ra với format là decision_tree_max_depth_{depth} với depth là độ sâu None, 2, 3, 4, 5, 6, 7.
- Bảng thống kệ độ chính xác của mô hình theo chiều sâu của cây quyết định.

max_depth	None	2	3	4	5	6	7
accuracy	0.8662	0.8861	0.8978	0.8974	0.8970	0.8954	0.8917

- Biểu đồ trực quan mối liên hệ giữa độ chính xác của dự đoán và chiều sâu của cây quyết định.



Nhận xét:

1. **Độ sâu tối ưu là 3**:
   Khi **max_depth** bằng **3**, mô hình đạt độ chính xác cao nhất, khoảng **0.898**. Đây là độ sâu phù hợp để mô hình hoạt động tốt nhất.
2. **Hiện tượng quá khớp (overfitting) khi tăng độ sâu**:
   Khi tăng **max_depth** lớn hơn **3** (4, 5, 6, 7), độ chính xác bắt đầu giảm dần. Điều này cho thấy mô hình càng phức tạp thì càng dễ xảy ra hiện tượng **quá khớp**, dẫn đến mô hình học cả nhiễu của dữ liệu thay vì khái quát hóa tốt.
3. **Hiện tượng dưới khớp (underfitting) tại độ sâu 2**:
   Khi max_depth = 2, độ chính xác thấp hơn đáng kể (khoảng **0.886**), cho thấy mô hình còn quá đơn giản và không đủ khả năng nắm bắt mối quan hệ trong dữ liệu.

6. Tài liệu tham khảo

[1] : https://machinelearningcoban.com/tabml_book/ch_model/decision_tree.html

[2] : https://viblo.asia/p/decision-tree-Do754bbBZM6

[3] : Moodle FIT, “Bài 7: Tham khảo 1” trong “Cơ sở trí tuệ nhân tạo”