

Câu 1: Cho CSDL sau và $\text{minsupp} = 60\%$ và $\text{minconf} = 100\%$

Đo TID bằng 1, 2, 3, 4, 5

TID	Items
10	D, H, C, A, B, K, M
20	E, H, D, G, P, I
30	B, C, D, G, H, K
40	E, A, C, B, P, I
50	K, B, M, F, H, D

Chuyển đổi bảng dl:	Item	Tidset	Item	Diffset
	A	1,4	B	2
	B	1,3,4,5	C	2,5
	C	1,3,4	D	4
	D	1,2,3,5	H	4
	E	2,4	K	2,4
	F	5		
	G	2,3		
	H	1,2,3,5		
	I	2,4		
	K	1,3,5		
	M	1,5		
	P	2,4		

Do $r = \frac{t(PXY)}{d(PXY)} \geq 1$
nên sử dụng diffset sẽ có lợi hơn

- a) Liệt kê các tập phô biến tối đa và tập phô biến đóng thỏa mãn ngưỡng minsupp đã cho sử dụng thuật toán Apriori.

Genmax

Charm

- b) Tìm các luật kết hợp có dạng sau và thỏa mãn ngưỡng minsupp, minconf đã cho sử dụng thuật toán Apriori

- item1 & item2 \rightarrow item3 & item4 (về trái và phải của luật đều có 2 hạng mục)
- D \rightarrow item (về phải có một hạng mục khác với hạng mục D)

Yêu cầu trình bày chi tiết các bước (không chỉ liệt kê tập luật tìm được)

Câu 2: Cho tập dữ liệu gồm 7 điểm trong không gian 2 chiều : P1, P2, P3, P4, P5, P6, P7. Cho ma trận khoảng cách giữa các điểm như trong bảng 1.

- a) Hãy sử dụng *lần lượt* thuật toán AGNES với **Single link** và **Complete link** để gom nhóm (**trình bày chi tiết các bước**). Vẽ sơ đồ hình cây (dendrogram) cho kết quả gom nhóm. (Sơ đồ hình cây phải vẽ rõ ràng để nhận biết được thứ tự và giá trị của vị trí các NHÓM gộp lại với nhau.)

- b) Dựa trên sơ đồ hình cây tương ứng (dùng Single Link/ Complete Link) xác định 3 nhóm thu được. So sánh kết quả.

Bảng 1 . Ma trận khoảng cách cho Câu 2

	P1	P2	P3	P4	P5	P6	P7
P1	0.00	0.27	0.23	0.56	0.17	0.40	0.14
P2	0.27	0.00	0.06	0.75	0.33	0.25	0.26
P3	0.23	0.06	0.00	0.59	0.28	0.24	0.22
P4	0.56	0.75	0.59	0.00	0.44	0.48	0.46
P5	0.17	0.33	0.28	0.44	0.00	0.37	0.09
P6	0.40	0.25	0.24	0.48	0.37	0.00	0.31
P7	0.14	0.26	0.22	0.46	0.09	0.31	0.00

Câu 3: Sử dụng **phương pháp cây quyết định** để tìm các luật phân lớp từ bảng dữ liệu sau. Giả sử thuộc tính “kết quả” là thuộc tính phân lớp.

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Nam	mưa
3	nhiều	trung bình	Bắc	mưa
4	ít	thấp	Bắc	không mưa
5	nhiều	thấp	Bắc	mưa
6	nhiều	cao	Bắc	mưa
7	nhiều	thấp	Nam	không mưa
8	ít	cao	Nam	không mưa

$$\text{Gain}(S, Mây) = 1 - \frac{3}{8} \left(\frac{-2}{3} \log_2 \frac{2}{3} \right) - \frac{5}{8} \left(-\frac{1}{5} \log_2 \frac{1}{5} - \frac{4}{5} \log_2 \frac{4}{5} \right) = 0.549$$

$$\text{Gain}(S, Gió) = 0.049$$

$$\text{Gain}(S, Áp suất) = 0.156$$

\Rightarrow Mây

$$\text{Entropy}(S, Mây) = 0$$

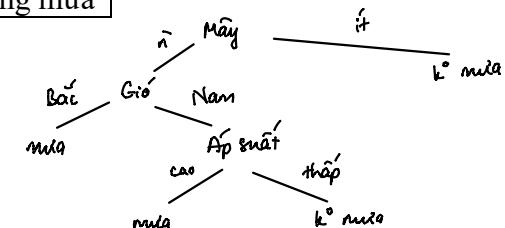
$$\text{Gain}(S_{\text{nhiều}}, Áp suất) = 0.722 - 0.4 = 0.322$$

$$\text{Gain}(S_{\text{nhiều}}, Gió) = 0.722 - 0.4 = 0.322$$

\Rightarrow Gió

$$\text{Gain}(S_{\text{nhiều}}, \text{Nam}, Áp suất) = 1 - 0 = 1$$

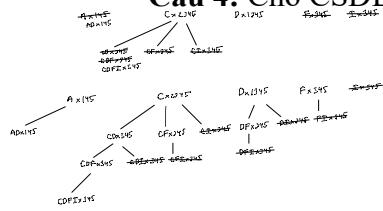
\Rightarrow Áp suất



	Support	C ₁	Support	C ₂	Support	C ₃	Support	C ₄
F	3			A, C	2	X	D, I	3
A	3			A, D	3		F, I	3
C	4			A, F	2	X	C, D, F	3
D	4			A, I	2	X	C, D, I	3
F	3			C, D	3		C, F, I	3
I	3			C, F	3		D, F, I	3

→ Tập pb đơn: C, D, AD
pb tối đa: CDFI, AD

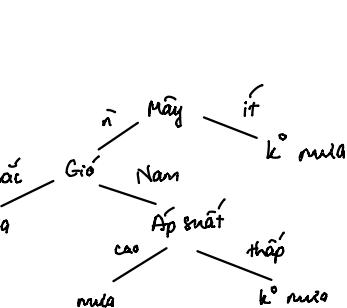
Câu 4: Cho CSDL sau



TID	A	B	C	D	E	F	G	H	I
10	1			1			1	1	
20			1		1				
30		1	1	1		1			1
40	1		1	1	1	1	1		1
50	1		1	1		1		1	1

- Hãy sử dụng **một** trong hai thuật toán : **Apriori** hoặc **FP-Growth** để tìm **tất cả** các tập phô biến thỏa mãn ngưỡng **minsupp=60%**. Liệt kê các tập phô biến tối đa và tập bao phô biến.
- Tìm các luật kết hợp được xây dựng từ tập phô biến tối đa, thỏa mãn ngưỡng **minconf =80%**.

Câu 5: Cho CSDL sau :



Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Nam	mưa
3	ít	thấp	Bắc	không mưa
4	nhiều	trung bình	Bắc	mưa
5	nhiều	thấp 0	Nam	không mưa
6	nhiều	thấp	Bắc	mưa
7	ít	cao	Nam	không mưa
8	nhiều	cao	Bắc	mưa

- Sử dụng **thuật toán ILA** để tìm các luật phân lớp với cột “**Kết quả**” là thuộc tính phân lớp. **Sử dụng bộ luật phân lớp tìm được để xác định lớp cho các đối tượng mới :**

Đối tượng	Mây	Áp suất	Gió	Kết quả
9	ít	trung bình	Bắc	?
10	ít	thấp	Nam	?
11	nhiều	trung bình	Nam	?

- L1: Nếu mây ít thì kết quả k' mưa
L2: Nếu áp suất thấp, gió nam thì k' mưa
L3: Nếu gió bắc thì k' mưa
L4: Nếu mây n, áp suất cao thì k' mưa
L5: Nếu mây n, áp suất thấp, gió bắc thì mưa

- Sử dụng thuật toán **cây quyết định** để tìm các luật phân lớp với cột “**Kết quả**” là thuộc tính phân lớp. **Sử dụng bộ luật phân lớp tìm được để xác định lớp cho các đối tượng mới ở trên và so sánh kết quả với câu a).** DT k' xác đc lớp cho mẫu 11

Câu 6: Cho CSDL sau :

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Bắc	mưa
3	ít	thấp	Bắc	không mưa
4	nhiều	thấp	Bắc	mưa
5	nhiều	trung bình	Bắc	mưa
6	ít	cao	Nam	không mưa
7	nhiều	cao	Nam	mưa
8	nhiều	thấp	Nam	không mưa

Sử dụng thuật toán Naïve Bayes để xác định lớp cho mẫu mới sau:

$$X_1: S(\text{"k' mưa"}) = P(\text{"k' mưa"}). R_1(\text{"ít", "k' mưa"}). R_2(\text{"thấp", "k' mưa"}). R_3(\text{"Nam", "k' mưa"}) \\ = \frac{4+1}{8+2} \cdot \frac{3+1}{4+2} \cdot \frac{2+1}{4+3} \cdot \frac{2+1}{4+2} = 0.071$$

$$S(\text{"có mưa"}) = P(\text{"có mưa"}). R_1(\dots) \dots \\ = \frac{4+1}{8+2} \cdot \frac{0+1}{4+2} \cdot \frac{1+1}{4+3} \cdot \frac{1+1}{4+2} = 0.008$$

$$X_2 : S("k^{\circ} mua") = \frac{4+1}{8+2} \cdot \frac{3+1}{4+2} \cdot \frac{0+1}{4+3} \cdot \frac{2+1}{4+2} = 0.024$$

$$S("có mua") = \frac{4+1}{8+2} \cdot \frac{0+1}{4+2} \cdot \frac{1+1}{4+3} \cdot \frac{3+1}{4+2} = 0.016$$

$$X_3 : S("k^{\circ} mua") = \frac{4+1}{8+2} \cdot \frac{1+1}{4+2} \cdot \frac{2+1}{4+3} \cdot \frac{2+1}{4+2} = 0.036$$

$$S("có mua") = \frac{4+1}{8+2} \cdot \frac{4+1}{4+2} \cdot \frac{2+1}{4+3} \cdot \frac{3+1}{4+2} = 0.119$$

$$X_4 : S("k^{\circ} mua") = \frac{4+1}{8+2} \cdot \frac{3+1}{4+2} \cdot \frac{0+1}{4+3} \cdot \frac{2+1}{4+2} = 0.024$$

$$S("có mua") = \frac{4+1}{8+2} \cdot \frac{0+1}{4+2} \cdot \frac{1+1}{4+3} \cdot \frac{1+1}{4+2} = 0.008$$

Đôi tượng	Mây	Áp suất	Gió	Kết quả
9	ít	thấp	Nam	?
10	ít	trung bình	Bắc	?
11	nhiều	cao	Bắc	?
12	nhiều	trung bình	Nam	?

k° mua
 k° mua
 có mua
 k° mua

Câu 7: Cho bảng dữ liệu thống kê kết quả của một thuật toán phân lớp số khách hàng đến siêu thị có mua hay không mua sản phẩm trong 1 tháng:

Lớp dự đoán				
Lớp thực	Lớp	Mua	Không mua	
	Mua	8986	1009	
	Không mua	1358	2547	

$$\text{confusion matrix} = \begin{bmatrix} 8986 & 1009 \\ 1358 & 2547 \end{bmatrix}$$

- Lập ma trận sai số (confusion matrix) $\text{Accuracy} = \frac{8986 + 2547}{8986 + 1009 + 1358 + 2547} = 82.97\%$

- Tính các độ đo accuracy, error rate, sensitivity, specificity, precision

$$\text{Precision} = \frac{8986}{8986 + 1358} = 86.87\%$$

$$\text{Error rate} = 1 - \text{accuracy}$$

$$= 1 - \frac{1009}{8986 + 1009} = 17.03\%$$

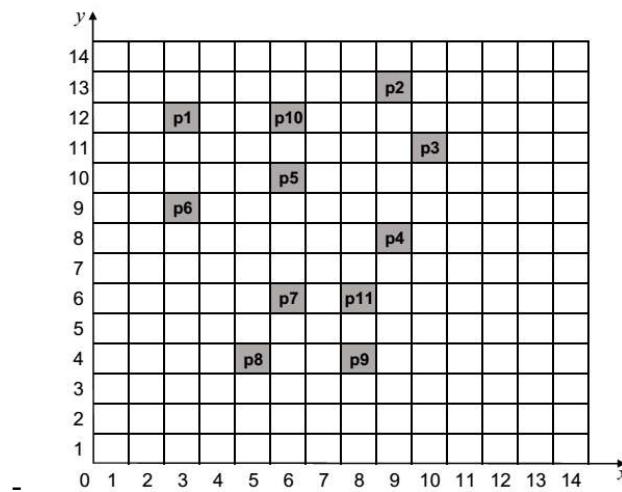
$$\text{Sensitivity} = 1 - \frac{1009}{1009 + 8986} = 89.9\%$$

$$\text{Specificity} = \frac{2547}{2547 + 1358} = 65.22\%$$

Câu 8: Cho các mẫu dữ liệu được phân bố trong không gian hai chiều Oxy như hình vẽ 1 (trang sau). Ví dụ: điểm P1 ở tọa độ (3,12). Giả sử người ta tiến hành gán nhãn cho mỗi điểm như sau:

p1:xanh, p2:xanh, p3:đỏ, p4:xanh, p5:đỏ, p6:xanh, p7:đỏ, p8:đỏ, p9:xanh.

Sử dụng thuật toán k-NN với khoảng cách Euclidean để phân lớp 2 mẫu sau: p10, p11 với số lân cận k = 3. Thực hiện việc tính toán đầy đủ.



- Hình 1: Phân bố các điểm dữ liệu trong không gian Oxy

Gợi ý: Công thức Euclidean của 2 điểm A, B trong không gian Oxy:

$$AB = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

Câu 9: Cho tập dữ liệu gồm 12 giá trị như bên dưới (đã sắp xếp theo thứ tự tăng dần).

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

a. Hãy áp dụng phương pháp chia giỏ để chia dữ liệu thành 3 giỏ bằng hai phương pháp:

- Chia giỏ theo độ rộng $\left[\frac{5}{q}, \frac{75}{q} \right] \quad \left[\frac{75}{2}, \frac{145}{2} \right] \quad \left[\frac{145}{2}, \frac{215}{2} \right]$
- Chia giỏ theo độ sâu $\left[\frac{5}{4}, \frac{18}{4} \right] \quad \left[\frac{15}{4}, \frac{55}{4} \right] \quad \left[\frac{72}{4}, \frac{215}{4} \right]$

b. Áp dụng làm tròn bằng giá trị trung bình, giá trị trung vị và biên giỏ cho trường hợp

chia giỏ theo độ sâu. Means: Bin 1: 9.75, 9.75, ... Bin 3: 145.75, ... Boundaries: Bin 1: 5, 13, 13, 13 Bin 3: 72, 72, 215, 215
Bin 2: 38.75, ... Bin 2: 15, 15, 55, 55

~~Câu 10:~~ Cho tập dữ liệu gồm 8 điểm trong không gian 2 chiều: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Hãy sử dụng lần lượt thuật toán **DBSCAN** để gom nhóm với Eps = 2 và Minpts = 2.

Solve

- [Phần 1: Tiền xử lý](#)
 - [1. Tiền xử lý Dữ liệu](#)
 - [**Bài tập 1**](#)
- [Phần 2: Khai thác Mẫu Phổ biến và Luật Kết hợp](#)
 - [2. Khai thác Tập Phổ biến \(Frequent Itemset Mining\)](#)
 - [**Bài tập 2**](#)
 - [**Bài tập 3**](#)
 - [4. Khai thác Luật Kết hợp \(Association Rule Mining\)](#)
 - [**Bài tập 4**](#)
 - [**Bài tập 5**](#)
- [Phần 3: Phân lớp \(Classification\)](#)
 - [5. Giới thiệu về Phân lớp & Cây Quyết định \(Decision Tree\)](#)
 - [**Bài tập 6**](#)
 - [**Bài tập 7**](#)
 - [6. Phân lớp Naïve Bayes](#)
 - [**Bài tập 8**](#)
 - [7. Phân lớp k-Nearest Neighbors \(k-NN\)](#)
 - [**Bài tập 9**](#)
 - [8. Phân lớp dựa trên Luật \(Rule-Based Classification - Tập trung vào ILA\)](#)
 - [**Bài tập 10**](#)
 - [9. Đánh giá Mô hình Phân lớp \(Model Evaluation\)](#)
 - [**Bài tập 11**](#)
- [Phần 4: Phân cụm \(Clustering\)](#)
 - [10. Phân cụm Phân cấp \(Hierarchical Clustering - AGNES\)](#)
 - [**Bài tập 12**](#)
 - [11. Phân cụm Mật độ \(Density-Based Clustering - DBSCAN\)](#)
 - [**Bài tập 13**](#)

Phần 1: Tiền xử lý

1. Tiền xử lý Dữ liệu

Bài tập 1

Cho tập dữ liệu gồm 12 giá trị như bên dưới (đã sắp xếp theo thứ tự tăng dần).

Dữ liệu: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Yêu cầu:

a) Chia giỏ (Binning)

Hãy áp dụng phương pháp chia giỏ để chia dữ liệu thành **3 giỏ** bằng hai phương pháp:

- Chia giỏ theo độ rộng (Equal Width)
- Chia giỏ theo độ sâu (Equal Depth / Equal Frequency)

b) Làm trơn (Smoothing)

Áp dụng làm trơn bằng **giá trị trung bình (mean)**, **giá trị trung vị (median)** và **biên giới (boundaries)** cho trường hợp chia giỏ theo độ sâu.

Giải:

Dữ liệu đã cho (đã sắp xếp): $D = \{5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215\}$.

Tổng số giá trị: $N = 12$.

Số giỏ cần chia: $k = 3$.

a) Chia giỏ (Binning)

- **Chia giỏ theo độ rộng (Equal Width):**

1. Xác định giá trị nhỏ nhất (min) và lớn nhất (max): $\min = 5$, $\max = 215$.
2. Tính độ rộng của mỗi giỏ:

$$\text{width} = \frac{\max - \min}{k} = \frac{215 - 5}{3} = \frac{210}{3} = 70$$

3. Xác định các khoảng của giỏ:

- Giỏ 1: $[5, 5 + 70] = [5, 75]$
- Giỏ 2: $[75, 75 + 70] = [75, 145]$
- Giỏ 3: $[145, 145 + 70] = [145, 215]$ (Giỏ cuối cùng chứa cả giá trị max)

4. Phân chia dữ liệu vào các giỏ:

- **Giỏ 1:** $\{5, 10, 11, 13, 15, 35, 50, 55, 72\}$
- **GiỎ 2:** $\{92\}$
- **GiỎ 3:** $\{204, 215\}$

- **Chia giỏ theo độ sâu (Equal Depth / Equal Frequency):** (*Thông thường, các giỏ đầu tiên sẽ được ưu tiên chứa nhiều phần tử hơn*)

1. Xác định số lượng phần tử trong mỗi giỏ:

$$\text{depth} = \frac{N}{k} = \frac{12}{3} = 4$$

2. Phân chia dữ liệu thành các giỏ, mỗi giỏ chứa 4 phần tử:

- **Giỏ 1:** {5, 10, 11, 13}
- **GiỎ 2:** {15, 35, 50, 55}
- **GiỎ 3:** {72, 92, 204, 215}

b) Làm trơn (Smoothing) cho trường hợp chia giỏ theo độ sâu

Áp dụng các phương pháp làm trơn cho các giỏ đã chia theo độ sâu:

- **Làm trơn bằng giá trị trung bình (Mean Smoothing):**
 - Giỏ 1: $\text{Mean}_1 = \frac{5+10+11+13}{4} = \frac{39}{4} = 9.75$. Giỏ mới: {9.75, 9.75, 9.75, 9.75}
 - Giỏ 2: $\text{Mean}_2 = \frac{15+35+50+55}{4} = \frac{155}{4} = 38.75$. Giỏ mới: {38.75, 38.75, 38.75, 38.75}
 - GiỎ 3: $\text{Mean}_3 = \frac{72+92+204+215}{4} = \frac{583}{4} = 145.75$. GiỎ mới: {145.75, 145.75, 145.75, 145.75}
- **Làm trơn bằng giá trị trung vị (Median Smoothing):**
 - Giỏ 1: {5, 10, 11, 13}. $\text{Median}_1 = \frac{10+11}{2} = 10.5$. Giỏ mới: {10.5, 10.5, 10.5, 10.5}
 - GiỎ 2: {15, 35, 50, 55}. $\text{Median}_2 = \frac{35+50}{2} = 42.5$. GiỎ mới: {42.5, 42.5, 42.5, 42.5}
 - GiỎ 3: {72, 92, 204, 215}. $\text{Median}_3 = \frac{92+204}{2} = \frac{296}{2} = 148$. GiỎ mới: {148, 148, 148, 148}
- **Làm trơn bằng biên giỏ (Bin Boundaries Smoothing):**
(Thay thế mỗi giá trị trong giỏ bằng biên gần nhất của giỏ đó)
 - GiỎ 1: {5, 10, 11, 13}. Biên: min = 5, max = 13.
 - 10 gần 13 hơn (khoảng cách 3 so với 5).
 - 11 gần 13 hơn (khoảng cách 2 so với 6).
 - GiỎ mới: {5, 13, 13, 13}
 - GiỎ 2: {15, 35, 50, 55}. Biên: min = 15, max = 55.
 - 35 cách đều 15 và 55 (khoảng cách 20). Chọn biên dưới: 15.
 - 50 gần 55 hơn (khoảng cách 5 so với 35).
 - GiỎ mới: {15, 15, 55, 55}
 - GiỎ 3: {72, 92, 204, 215}. Biên: min = 72, max = 215.
 - 92 gần 72 hơn (khoảng cách 20 so với 123).
 - 204 gần 215 hơn (khoảng cách 11 so với 132).
 - GiỎ mới: {72, 72, 215, 215}

Phần 2: Khai thác Mẫu Phổ biến và Luật Kết hợp

2. Khai thác Tập Phổ biến (Frequent Itemset Mining)

Bài tập 2

Cho Cơ sở dữ liệu (CSDL) sau và $\text{minsupp} = 60\%$.

CSDL Giao dịch:

TID	Items
10	D, H, C, A, B, K, M
20	E, H, D, G, P, I
30	B, C, D, G, H, K
40	E, A, C, B, P, I
50	K, B, M, F, H, D

Yêu cầu:

Sử dụng thuật toán Apriori: Liệt kê các tập phỗ biến tối đại (maximal frequent itemsets) và tập phỗ biến đóng (closed frequent itemsets) thoả mãn ngưỡng minsupp đã cho.

Yêu cầu trình bày chi tiết các bước.

Giải:

1. Xác định ngưỡng hỗ trợ tuyệt đối:

CSDL có 5 giao dịch ($|D| = 5$).

Ngưỡng hỗ trợ tối thiểu tương đối $\text{minsupp} = 60\%$.

Ngưỡng hỗ trợ tối thiểu tuyệt đối $\text{minsup_count} = 5 \times 60\% = 3$.

2. Áp dụng Thuật toán Apriori:

- Bước 1: Tìm tập phỗ biến 1-itemset (L_1)**

Quét CSDL lần 1 để đếm support cho từng item:

- $\text{Sup}(A) = 2 (< 3)$
- $\text{Sup}(B) = 4 (\geq 3)$
- $\text{Sup}(C) = 3 (\geq 3)$
- $\text{Sup}(D) = 4 (\geq 3)$
- $\text{Sup}(E) = 2 (< 3)$
- $\text{Sup}(F) = 1 (< 3)$
- $\text{Sup}(G) = 2 (< 3)$
- $\text{Sup}(H) = 4 (\geq 3)$
- $\text{Sup}(I) = 2 (< 3)$

- $\text{Sup}(K) = 3 (\geq 3)$
- $\text{Sup}(M) = 2 (< 3)$
- $\text{Sup}(P) = 2 (< 3)$

$$L_1 = \{\{B\} : 4, \{C\} : 3, \{D\} : 4, \{H\} : 4, \{K\} : 3\}$$

- **Bước 2: Tìm tập phỏ biến 2-itemset (L_2)**

- Sinh ứng viên C_2 từ $L_1 \bowtie L_1$:

$$C_2 = \{\{B,C\}, \{B,D\}, \{B,H\}, \{B,K\}, \{C,D\}, \{C,H\}, \{C,K\}, \{D,H\}, \{D,K\}, \{H,K\}\}$$

(Không có ứng viên nào bị tia ở bước này vì tất cả tập con 1-itemset đều thuộc L_1)

- Quét CSDL lần 2 để đếm support cho C_2 :

$$\text{Sup}(\{B,C\}) = |\{10, 30, 40\}| = 3 (\geq 3)$$

$$\text{Sup}(\{B,D\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$\text{Sup}(\{B,H\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$\text{Sup}(\{B,K\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$\text{Sup}(\{C,D\}) = |\{10, 30\}| = 2 (< 3)$$

$$\text{Sup}(\{C,H\}) = |\{10, 30\}| = 2 (< 3)$$

$$\text{Sup}(\{C,K\}) = |\{10, 30\}| = 2 (< 3)$$

$$\text{Sup}(\{D,H\}) = |\{10, 20, 30, 50\}| = 4 (\geq 3)$$

$$\text{Sup}(\{D,K\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$\text{Sup}(\{H,K\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$L_2 = \{\{B,C\} : 3, \{B,D\} : 3, \{B,H\} : 3, \{B,K\} : 3, \{D,H\} : 4, \{D,K\} : 3, \{H,K\} : 3\}$$

- **Bước 3: Tìm tập phỏ biến 3-itemset (L_3)**

- Sinh ứng viên C_3 từ $L_2 \bowtie L_2$:

- $\{B,D\}$ và $\{B,H\} \rightarrow \{B,D,H\}$ (Tập con $\{B,D\}$, $\{B,H\}$, $\{D,H\} \in L_2$)

- $\{B,D\}$ và $\{B,K\} \rightarrow \{B,D,K\}$ (Tập con $\{B,D\}$, $\{B,K\}$, $\{D,K\} \in L_2$)

- $\{B,H\}$ và $\{B,K\} \rightarrow \{B,H,K\}$ (Tập con $\{B,H\}$, $\{B,K\}$, $\{H,K\} \in L_2$)

- $\{D,H\}$ và $\{D,K\} \rightarrow \{D,H,K\}$ (Tập con $\{D,H\}$, $\{D,K\}$, $\{H,K\} \in L_2$)

(Các kết hợp khác không thỏa mãn điều kiện join hoặc bị tia do có tập con không thuộc L_2 . Ví dụ: $\{B,C\}$ và $\{B,D\} \rightarrow \{B,C,D\}$, tập con $\{C,D\} \notin L_2 \rightarrow$ Loại)

$$C_3 = \{\{B,D,H\}, \{B,D,K\}, \{B,H,K\}, \{D,H,K\}\}$$

- Quét CSDL lần 3 để đếm support cho C_3 :

$$\text{Sup}(\{B,D,H\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$\text{Sup}(\{B,D,K\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$\text{Sup}(\{B,H,K\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$\text{Sup}(\{D,H,K\}) = |\{10, 30, 50\}| = 3 (\geq 3)$$

$$L_3 = \{\{B,D,H\} : 3, \{B,D,K\} : 3, \{B,H,K\} : 3, \{D,H,K\} : 3\}$$

- **Bước 4: Tìm tập phỏ biến 4-itemset (L_4)**

- Sinh ứng viên C_4 từ $L_3 \bowtie L_3$:

- $\{B,D,H\}$ và $\{B,D,K\} \rightarrow \{B,D,H,K\}$ (Tập con $\{B,D,H\}$, $\{B,D,K\}$, $\{B,H,K\}$, $\{D,H,K\}$ đều $\in L_3$)

$$C_4 = \{\{B, D, H, K\}\}$$

- Quét CSDL lần 4 để đếm support cho C_4 :

* $\text{Sup}(\{B, D, H, K\}) = |\{10, 30, 50\}| = 3 (\geq 3)$

$$L_4 = \{\{B, D, H, K\} : 3\}$$

- **Bước 5: Tìm tập phỗ biến 5-itemset (L_5)**

- Không thể sinh C_5 từ L_4 . Dừng.

3. Tập hợp tất cả các tập phỗ biến ($L = L_1 \cup L_2 \cup L_3 \cup L_4$):

- **Sup=4:** $\{B\}, \{D\}, \{H\}, \{D, H\}$
- **Sup=3:** $\{C\}, \{K\}, \{B, C\}, \{B, D\}, \{B, H\}, \{B, K\}, \{D, K\}, \{H, K\}, \{B, D, H\}, \{B, D, K\}, \{B, H, K\}, \{D, H, K\}, \{B, D, H, K\}$

4. Xác định Tập phỗ biến đóng (Closed Frequent Itemsets - CFIs):

Là tập phỗ biến X mà không tồn tại tập cha thực sự $Y \supset X$ có $\text{sup}(Y) = \text{sup}(X)$.

- $\{B\}$ (Sup=4) không đóng vì $\text{sup}(\{D, H\}) = 4$.
- $\{D\}$ (Sup=4) không đóng vì $\text{sup}(\{D, H\}) = 4$.
- $\{H\}$ (Sup=4) không đóng vì $\text{sup}(\{D, H\}) = 4$.
- $\{C\}$ (Sup=3) không đóng vì $\text{sup}(\{B, C\}) = 3$.
- $\{K\}$ (Sup=3) không đóng vì $\text{sup}(\{B, K\}) = 3$ (hoặc $\{D, K\}, \{H, K\}$).
- $\{B, C\}$ (Sup=3) là đóng (không có cha nào sup=3).
- $\{B, D\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, H\}) = 3$.
- $\{B, H\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, H\}) = 3$.
- $\{B, K\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, K\}) = 3$.
- $\{D, H\}$ (Sup=4) là đóng (không có cha nào sup=4).
- $\{D, K\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, K\}) = 3$.
- $\{H, K\}$ (Sup=3) không đóng vì $\text{sup}(\{B, H, K\}) = 3$.
- $\{B, D, H\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, H, K\}) = 3$.
- $\{B, D, K\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, H, K\}) = 3$.
- $\{B, H, K\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, H, K\}) = 3$.
- $\{D, H, K\}$ (Sup=3) không đóng vì $\text{sup}(\{B, D, H, K\}) = 3$.
- $\{B, D, H, K\}$ (Sup=3) là đóng (không có cha nào).

Kết quả CFIs: $\{\{B, C\}, \{D, H\}, \{B, D, H, K\}\}$

5. Xác định Tập phỗ biến tối đại (Maximal Frequent Itemsets - MFIs):

Là tập phỗ biến X mà không tồn tại tập cha thực sự $Y \supset X$ cũng là tập phỗ biến.

- $\{B, D, H, K\}$ là MFI vì không có tập cha nào trong L .

- Các tập con của {B,D,H,K} không phải MFI.
- {B,C} là MFI vì không có tập cha nào của nó trong L .

Kết quả MFI: $\{\{B, C\}, \{B, D, H, K\}\}$

Bài tập 3

Cho CSDL sau (dạng nhị phân):

TID	A	B	C	D	E	F	G	H	I
10	1			1			1	1	
20				1		1			
30		1	1	1			1		1
40	1			1	1	1	1		1
50	1			1	1		1	1	

Yêu cầu:

Hãy sử dụng một trong hai thuật toán: **Apriori** hoặc **FP-Growth** để tìm tất cả các tập phỗ biến thỏa mãn ngưỡng $\text{minsupp}=60\%$. Liệt kê các tập phỗ biến tối đại và tập bao phỗ biến (closed frequent itemsets).

Giải:

1. Xác định ngưỡng hỗ trợ tuyệt đối:

CSDL có 5 giao dịch ($|D| = 5$).

Ngưỡng hỗ trợ tối thiểu tương đối $\text{minsupp} = 60\%$.

Ngưỡng hỗ trợ tối thiểu tuyệt đối $\text{minsup_count} = 5 \times 60\% = 3$.

2. Áp dụng Thuật toán Apriori: (Có thể dùng FP-Growth, nhưng Apriori được chọn ở đây)

- **Bước 1: Tìm tập phỗ biến 1-itemset (L_1)**

Quét CSDL lần 1:

- $\text{Sup}(A) = |\{10, 40, 50\}| = 3 (\geq 3)$
- $\text{Sup}(B) = |\{30\}| = 1 (< 3)$
- $\text{Sup}(C) = |\{20, 30, 40, 50\}| = 4 (\geq 3)$
- $\text{Sup}(D) = |\{10, 30, 40, 50\}| = 4 (\geq 3)$
- $\text{Sup}(E) = |\{20, 40\}| = 2 (< 3)$
- $\text{Sup}(F) = |\{30, 40, 50\}| = 3 (\geq 3)$

- $\text{Sup}(G) = |\{10, 40\}| = 2 (< 3)$
 - $\text{Sup}(H) = |\{10, 50\}| = 2 (< 3)$
 - $\text{Sup}(I) = |\{30, 40, 50\}| = 3 (\geq 3)$
- $L_1 = \{\{A\} : 3, \{C\} : 4, \{D\} : 4, \{F\} : 3, \{I\} : 3\}$

- **Bước 2: Tìm tập phỏ biến 2-itemset (L_2)**

- Sinh ứng viên C_2 từ $L_1 \bowtie L_1$:
- $C_2 = \{\{A, C\}, \{A, D\}, \{A, F\}, \{A, I\}, \{C, D\}, \{C, F\}, \{C, I\}, \{D, F\}, \{D, I\}, \{F, I\}\}$
- Quét CSDL lần 2 để đếm support cho C_2 :
- $\text{Sup}(\{A, C\}) = |\{40, 50\}| = 2 (< 3)$
 $\text{Sup}(\{A, D\}) = |\{10, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{A, F\}) = |\{40, 50\}| = 2 (< 3)$
 $\text{Sup}(\{A, I\}) = |\{40, 50\}| = 2 (< 3)$
 $\text{Sup}(\{C, D\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{C, F\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{C, I\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{D, F\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{D, I\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{F, I\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
- $L_2 = \{\{A, D\} : 3, \{C, D\} : 3, \{C, F\} : 3, \{C, I\} : 3, \{D, F\} : 3, \{D, I\} : 3, \{F, I\} : 3\}$

- **Bước 3: Tìm tập phỏ biến 3-itemset (L_3)**

- Sinh ứng viên C_3 từ $L_2 \bowtie L_2$:
 - $\{C, D\}$ và $\{C, F\} \rightarrow \{C, D, F\}$ (Tập con $\{C, D\}$, $\{C, F\}$, $\{D, F\} \in L_2$)
 - $\{C, D\}$ và $\{C, I\} \rightarrow \{C, D, I\}$ (Tập con $\{C, D\}$, $\{C, I\}$, $\{D, I\} \in L_2$)
 - $\{C, F\}$ và $\{C, I\} \rightarrow \{C, F, I\}$ (Tập con $\{C, F\}$, $\{C, I\}$, $\{F, I\} \in L_2$)
 - $\{D, F\}$ và $\{D, I\} \rightarrow \{D, F, I\}$ (Tập con $\{D, F\}$, $\{D, I\}$, $\{F, I\} \in L_2$)

(Các kết hợp khác không thỏa mãn điều kiện join hoặc bị tẩy do có tập con không thuộc L_2 . Ví dụ: $\{A, D\}$ và $\{C, D\} \rightarrow \{A, C, D\}$, tập con $\{A, C\} \notin L_2 \rightarrow$ Loại)

$C_3 = \{\{C, D, F\}, \{C, D, I\}, \{C, F, I\}, \{D, F, I\}\}$
- Quét CSDL lần 3 để đếm support cho C_3 :

$\text{Sup}(\{C, D, F\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{C, D, I\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{C, F, I\}) = |\{30, 40, 50\}| = 3 (\geq 3)$
 $\text{Sup}(\{D, F, I\}) = |\{30, 40, 50\}| = 3 (\geq 3)$

$L_3 = \{\{C, D, F\} : 3, \{C, D, I\} : 3, \{C, F, I\} : 3, \{D, F, I\} : 3\}$

- **Bước 4: Tìm tập phỏ biến 4-itemset (L_4)**

- Sinh ứng viên C_4 từ $L_3 \bowtie L_3$:
 - $\{C, D, F\}$ và $\{C, D, I\} \rightarrow \{C, D, F, I\}$ (Tập con $\{C, D, F\}$, $\{C, D, I\}$, $\{C, F, I\}$, $\{D, F, I\}$ đều $\in L_3$)

$C_4 = \{\{C, D, F, I\}\}$

- Quét CSDL lần 4 để đếm support cho C_4 :
 - * $\text{Sup}(\{C,D,F,I\}) = |\{30, 40, 50\}| = 3 \geq 3$
 - $L_4 = \{\{C, D, F, I\} : 3\}$
- **Bước 5: Tìm tập phỗ biến 5-itemset (L_5)**
 - Không thể sinh C_5 từ L_4 . Dừng.

3. Tập hợp tất cả các tập phỗ biến ($L = L_1 \cup L_2 \cup L_3 \cup L_4$):

- **Sup=4:** $\{C\}, \{D\}$
- **Sup=3:** $\{A\}, \{F\}, \{I\}, \{A, D\}, \{C, D\}, \{C, F\}, \{C, I\}, \{D, F\}, \{D, I\}, \{F, I\}, \{C, D, F\}, \{C, D, I\}, \{C, F, I\}, \{D, F, I\}, \{C, D, F, I\}$

4. Xác định Tập phỗ biến đóng (Closed Frequent Itemsets - CFIs):

Là tập phỗ biến X mà không tồn tại tập cha thực sự $Y \supset X$ có $\text{sup}(Y) = \text{sup}(X)$.

- $\{C\}$ (Sup=4) là đóng.
- $\{D\}$ (Sup=4) là đóng.
- $\{A\}$ (Sup=3) không đóng vì $\text{sup}(\{A, D\}) = 3$.
- $\{F\}$ (Sup=3) không đóng vì $\text{sup}(\{C, F\}) = 3$.
- $\{I\}$ (Sup=3) không đóng vì $\text{sup}(\{C, I\}) = 3$.
- $\{A, D\}$ (Sup=3) là đóng.
- $\{C, D\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, F\}) = 3$.
- $\{C, F\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, F\}) = 3$.
- $\{C, I\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, I\}) = 3$.
- $\{D, F\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, F\}) = 3$.
- $\{D, I\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, I\}) = 3$.
- $\{F, I\}$ (Sup=3) không đóng vì $\text{sup}(\{C, F, I\}) = 3$.
- $\{C, D, F\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, F, I\}) = 3$.
- $\{C, D, I\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, F, I\}) = 3$.
- $\{C, F, I\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, F, I\}) = 3$.
- $\{D, F, I\}$ (Sup=3) không đóng vì $\text{sup}(\{C, D, F, I\}) = 3$.
- $\{C, D, F, I\}$ (Sup=3) là đóng.

Kết quả CFIs: $\{\{C\}, \{D\}, \{A, D\}, \{C, D, F, I\}\}$

5. Xác định Tập phỗ biến tối đại (Maximal Frequent Itemsets - MFIs):

Là tập phỗ biến X mà không tồn tại tập cha thực sự $Y \supset X$ cũng là tập phỗ biến.

- $\{C, D, F, I\}$ là MFI.
- Các tập con của $\{C, D, F, I\}$ không phải MFI.

- $\{A, D\}$ là MFI (không có tập cha nào trong L).

Kết quả MFI: $\{\{A, D\}, \{C, D, F, I\}\}$

4. Khai thác Luật Kết hợp (Association Rule Mining)

Bài tập 4

Sử dụng CSDL và `minsupp = 60%` từ **Bài tập 2**. Cho `minconf = 100%`.

Yêu cầu:

Sử dụng thuật toán Apriori: Tìm các luật kết hợp có dạng sau và thỏa mãn ngưỡng `minsupp`, `minconf` đã cho:

- `item1 & item2 -> item3 & item4` (về trái và phải của luật đều có 2 hạng mục)
- `D -> item` (về phải có một hạng mục khác với hạng mục D)

Yêu cầu trình bày chi tiết các bước.

Giải:

Dựa vào kết quả các tập phỏ biến tìm được ở Bài tập 2 với `minsup_count = 3` và yêu cầu `minconf = 100%`.

Các tập phỏ biến liên quan:

- $sup(\{B\}) = 4$
- $sup(\{D\}) = 4$
- $sup(\{H\}) = 4$
- $sup(\{K\}) = 3$
- $sup(\{B, D\}) = 3$
- $sup(\{B, H\}) = 3$
- $sup(\{B, K\}) = 3$
- $sup(\{D, H\}) = 4$
- $sup(\{D, K\}) = 3$
- $sup(\{H, K\}) = 3$
- $sup(\{B, D, H, K\}) = 3$

1. Tìm luật dạng `item1 & item2 -> item3 & item4`:

Luật này chỉ có thể được sinh ra từ tập phỏ biến 4-itemset $\{B, D, H, K\}$ (có $sup = 3 \geq 3$).

Ta cần kiểm tra các luật có thể sinh ra từ tập này với $conf \geq 100\%$.

Công thức: $conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} = \frac{sup(\{B,D,H,K\})}{sup(\{X\})}$

- Luật $\{B, D\} \rightarrow \{H, K\}$:
 $conf = \frac{sup(\{B,D,H,K\})}{sup(\{B,D\})} = \frac{3}{3} = 100\%. (\text{Thỏa mãn})$
- Luật $\{B, H\} \rightarrow \{D, K\}$:
 $conf = \frac{sup(\{B,D,H,K\})}{sup(\{B,H\})} = \frac{3}{3} = 100\%. (\text{Thỏa mãn})$
- Luật $\{B, K\} \rightarrow \{D, H\}$:
 $conf = \frac{sup(\{B,D,H,K\})}{sup(\{B,K\})} = \frac{3}{3} = 100\%. (\text{Thỏa mãn})$
- Luật $\{D, H\} \rightarrow \{B, K\}$:
 $conf = \frac{sup(\{B,D,H,K\})}{sup(\{D,H\})} = \frac{3}{4} = 75\%. (\text{Loại})$
- Luật $\{D, K\} \rightarrow \{B, H\}$:
 $conf = \frac{sup(\{B,D,H,K\})}{sup(\{D,K\})} = \frac{3}{3} = 100\%. (\text{Thỏa mãn})$
- Luật $\{H, K\} \rightarrow \{B, D\}$:
 $conf = \frac{sup(\{B,D,H,K\})}{sup(\{H,K\})} = \frac{3}{3} = 100\%. (\text{Thỏa mãn})$

Kết quả luật dạng 1: $\{B, D\} \rightarrow \{H, K\}$, $\{B, H\} \rightarrow \{D, K\}$, $\{B, K\} \rightarrow \{D, H\}$, $\{D, K\} \rightarrow \{B, H\}$, $\{H, K\} \rightarrow \{B, D\}$.

2. Tìm luật dạng $D \rightarrow \text{item}$ (với item khác D):

Ta cần kiểm tra các luật $D \rightarrow X$ sao cho $\{D, X\}$ là tập phỗ biến và $conf \geq 100\%$.

Công thức: $conf(D \rightarrow X) = \frac{sup(\{D,X\})}{sup(\{D\})}$

- Luật $D \rightarrow B$:
 $conf = \frac{sup(\{D,B\})}{sup(\{D\})} = \frac{3}{4} = 75\%. (\text{Loại})$
- Luật $D \rightarrow C$: $sup(\{D,C\}) = 2 < 3$. Tập $\{D, C\}$ không phỗ biến. (Loại)
- Luật $D \rightarrow H$:
 $conf = \frac{sup(\{D,H\})}{sup(\{D\})} = \frac{4}{4} = 100\%. (\text{Thỏa mãn})$
- Luật $D \rightarrow K$:
 $conf = \frac{sup(\{D,K\})}{sup(\{D\})} = \frac{3}{4} = 75\%. (\text{Loại})$

Kết quả luật dạng 2: $D \rightarrow H$.

Tổng kết các luật thỏa mãn yêu cầu:

- $\{B, D\} \rightarrow \{H, K\}$
- $\{B, H\} \rightarrow \{D, K\}$
- $\{B, K\} \rightarrow \{D, H\}$
- $\{D, K\} \rightarrow \{B, H\}$
- $\{H, K\} \rightarrow \{B, D\}$

- $D \rightarrow H$

Bài tập 5

Sử dụng CSDL và `minsupp=60%` từ **Bài tập 3**. Cho `minconf=80%`.

Yêu cầu:

Tìm các luật kết hợp được xây dựng từ tập phỏ biến tối đại (tìm được ở Bài tập 3), thỏa mãn `ngưỡng minconf=80%`.

Giải:

Dựa vào kết quả các tập phỏ biến tối đại (MFI) tìm được ở Bài tập 3 và yêu cầu `minconf = 80%`.

Các MFI là: $\{\{A, D\}, \{C, D, F, I\}\}$.

`minsup_count = 3`.

Các support count liên quan từ Bài tập 3:

- $sup(\{A\}) = 3$
- $sup(\{C\}) = 4$
- $sup(\{D\}) = 4$
- $sup(\{F\}) = 3$
- $sup(\{I\}) = 3$
- $sup(\{A, D\}) = 3$
- $sup(\{C, D\}) = 3$
- $sup(\{C, F\}) = 3$
- $sup(\{C, I\}) = 3$
- $sup(\{D, F\}) = 3$
- $sup(\{D, I\}) = 3$
- $sup(\{F, I\}) = 3$
- $sup(\{C, D, F\}) = 3$
- $sup(\{C, D, I\}) = 3$
- $sup(\{C, F, I\}) = 3$
- $sup(\{D, F, I\}) = 3$
- $sup(\{C, D, F, I\}) = 3$

1. Xét MFI $\{A, D\}$:

- Luật $A \rightarrow D$:

$$conf = \frac{sup(\{A,D\})}{sup(\{A\})} = \frac{3}{3} = 100\%. (\text{Thỏa mãn})$$

- Luật $D \rightarrow A$:

$$conf = \frac{sup(\{A,D\})}{sup(\{D\})} = \frac{3}{4} = 75\%. (\text{Loại})$$

2. Xét MFI $\{C, D, F, I\}$:

Tập phổ biến $\{C, D, F, I\}$ có $sup = 3$. Ta sinh các luật $X \rightarrow Y$ sao cho $X \cup Y = \{C, D, F, I\}$ và $X \cap Y = \emptyset$.

Công thức: $conf(X \rightarrow Y) = \frac{sup(\{C,D,F,I\})}{sup(X)} = \frac{3}{sup(X)}$.

Để $conf \geq 80\%$, ta cần $sup(X) \leq \frac{3}{0.8} = 3.75$. Vậy $sup(X)$ chỉ có thể là 3.

Kiểm tra các tập con X của $\{C, D, F, I\}$ có $sup(X) = 3$:

- **Tập con kích thước 1:** $\{A\}$ ($sup = 3$, không phải con của $\{C, D, F, I\}$), $\{F\}$ ($sup = 3$), $\{I\}$ ($sup = 3$)).
- **Tập con kích thước 2:** $\{A, D\}$ ($sup = 3$), $\{C, D\}$ ($sup = 3$), $\{C, F\}$ ($sup = 3$), $\{C, I\}$ ($sup = 3$), $\{D, F\}$ ($sup = 3$), $\{D, I\}$ ($sup = 3$), $\{F, I\}$ ($sup = 3$)).
- **Tập con kích thước 3:** $\{C, D, F\}$ ($sup = 3$), $\{C, D, I\}$ ($sup = 3$), $\{C, F, I\}$ ($sup = 3$), $\{D, F, I\}$ ($sup = 3$)).

Sinh luật từ các tập con X có $sup(X) = 3$:

- $F \rightarrow \{C, D, I\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $I \rightarrow \{C, D, F\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{C, D\} \rightarrow \{F, I\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{C, F\} \rightarrow \{D, I\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{C, I\} \rightarrow \{D, F\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{D, F\} \rightarrow \{C, I\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{D, I\} \rightarrow \{C, F\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{F, I\} \rightarrow \{C, D\}$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{C, D, F\} \rightarrow I$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{C, D, I\} \rightarrow F$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{C, F, I\} \rightarrow D$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)
- $\{D, F, I\} \rightarrow C$ ($conf = 3/3 = 100\%$). (**Thỏa mãn**)

Lưu ý: Các tập con có $sup = 4$ (như $\{C\}$, $\{D\}$) sẽ không tạo ra luật thỏa mãn $minconf = 80\%$ vì $3/4 = 75\% < 80\%$.

Tổng kết các luật thỏa mãn yêu cầu:

- Từ MFI $\{A, D\}$:

- $A \rightarrow D$
- Từ MFI {C, D, F, I}:
 - $F \rightarrow \{C, D, I\}$
 - $I \rightarrow \{C, D, F\}$
 - $\{C, D\} \rightarrow \{F, I\}$
 - $\{C, F\} \rightarrow \{D, I\}$
 - $\{C, I\} \rightarrow \{D, F\}$
 - $\{D, F\} \rightarrow \{C, I\}$
 - $\{D, I\} \rightarrow \{C, F\}$
 - $\{F, I\} \rightarrow \{C, D\}$
 - $\{C, D, F\} \rightarrow I$
 - $\{C, D, I\} \rightarrow F$
 - $\{C, F, I\} \rightarrow D$
 - $\{D, F, I\} \rightarrow C$

Phần 3: Phân lớp (Classification)

5. Giới thiệu về Phân lớp & Cây Quyết định (Decision Tree)

Bài tập 6

Cho CSDL sau

Bảng dữ liệu:

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Nam	mưa
3	nhiều	trung bình	Bắc	mưa
4	ít	thấp	Bắc	không mưa
5	nhiều	thấp	Bắc	mưa
6	nhiều	cao	Bắc	mưa
7	nhiều	thấp	Nam	không mưa
8	ít	cao	Nam	không mưa

Yêu cầu:

Sử dụng phương pháp **cây quyết định** để tìm các luật phân lớp từ bảng dữ liệu sau. Giả sử thuộc tính "Kết quả" là thuộc tính phân lớp.

Giải:

1. Tính toán độ lợi thông tin (Information Gain) để chọn thuộc tính gốc:

- **Tập dữ liệu gốc (S):** 8 đối tượng, 4 'mưa' (M), 4 'không mưa' (KM).
- **Entropy của tập gốc:**

$$Entropy(S) = -\frac{4}{8} \log_2 \left(\frac{4}{8} \right) - \frac{4}{8} \log_2 \left(\frac{4}{8} \right) = 1$$

- **Tính Gain cho từng thuộc tính:**

- **Thuộc tính 'Mây':**

- Giá trị 'ít' (3 đối tượng {1, 4, 8}): 0 M, 3 KM. $Entropy(Mây=ít) = 0$.
 - Giá trị 'nhiều' (5 đối tượng {2, 3, 5, 6, 7}): 4 M, 1 KM.

$$Entropy(Mây=nhiều) = -\frac{4}{5} \log_2 \left(\frac{4}{5} \right) - \frac{1}{5} \log_2 \left(\frac{1}{5} \right) \approx 0.7219.$$

- $Gain(Mây) = Entropy(S) - \left(\frac{3}{8} Entropy(Mây=ít) + \frac{5}{8} Entropy(Mây=nhiều) \right)$
 - $Gain(Mây) \approx 1 - \left(\frac{3}{8} \times 0 + \frac{5}{8} \times 0.7219 \right) \approx 1 - 0.4512 = 0.5488$.

- **Thuộc tính 'Áp suất':**

- Giá trị 'cao' (3 đối tượng {1, 6, 8}): 1 M, 2 KM.

$$Entropy(\text{Áp suất}=cao) = -\frac{1}{3} \log_2 \left(\frac{1}{3} \right) - \frac{2}{3} \log_2 \left(\frac{2}{3} \right) \approx 0.9183.$$

- Giá trị 'trung bình' (1 đối tượng {3}): 1 M, 0 KM. $Entropy(\text{Áp suất}=tb) = 0$.

- Giá trị 'thấp' (4 đối tượng {2, 4, 5, 7}): 2 M, 2 KM.

$$Entropy(\text{Áp suất}=tháp) = -\frac{2}{4} \log_2 \left(\frac{2}{4} \right) - \frac{2}{4} \log_2 \left(\frac{2}{4} \right) = 1.$$

- $Gain(\text{Áp suất}) = Entropy(S) - \left(\frac{3}{8} E(\text{cao}) + \frac{1}{8} E(\text{tb}) + \frac{4}{8} E(\text{tháp}) \right)$

$$Gain(\text{Áp suất}) \approx 1 - \left(\frac{3}{8} \times 0.9183 + \frac{1}{8} \times 0 + \frac{4}{8} \times 1 \right) \approx 1 - (0.3444 + 0 + 0.5) = 1 - 0.84$$

- **Thuộc tính 'Gió':**

- Giá trị 'Bắc' (5 đối tượng {1, 3, 4, 5, 6}): 3 M, 2 KM.

$$Entropy(\text{Gió=Bắc}) = -\frac{3}{5} \log_2 \left(\frac{3}{5} \right) - \frac{2}{5} \log_2 \left(\frac{2}{5} \right) \approx 0.9710.$$

- Giá trị 'Nam' (3 đối tượng {2, 7, 8}): 1 M, 2 KM.

$$Entropy(\text{Gió=Nam}) = -\frac{1}{3} \log_2 \left(\frac{1}{3} \right) - \frac{2}{3} \log_2 \left(\frac{2}{3} \right) \approx 0.9183.$$

- $Gain(\text{Gió}) = Entropy(S) - \left(\frac{5}{8} E(\text{Bắc}) + \frac{3}{8} E(\text{Nam}) \right)$

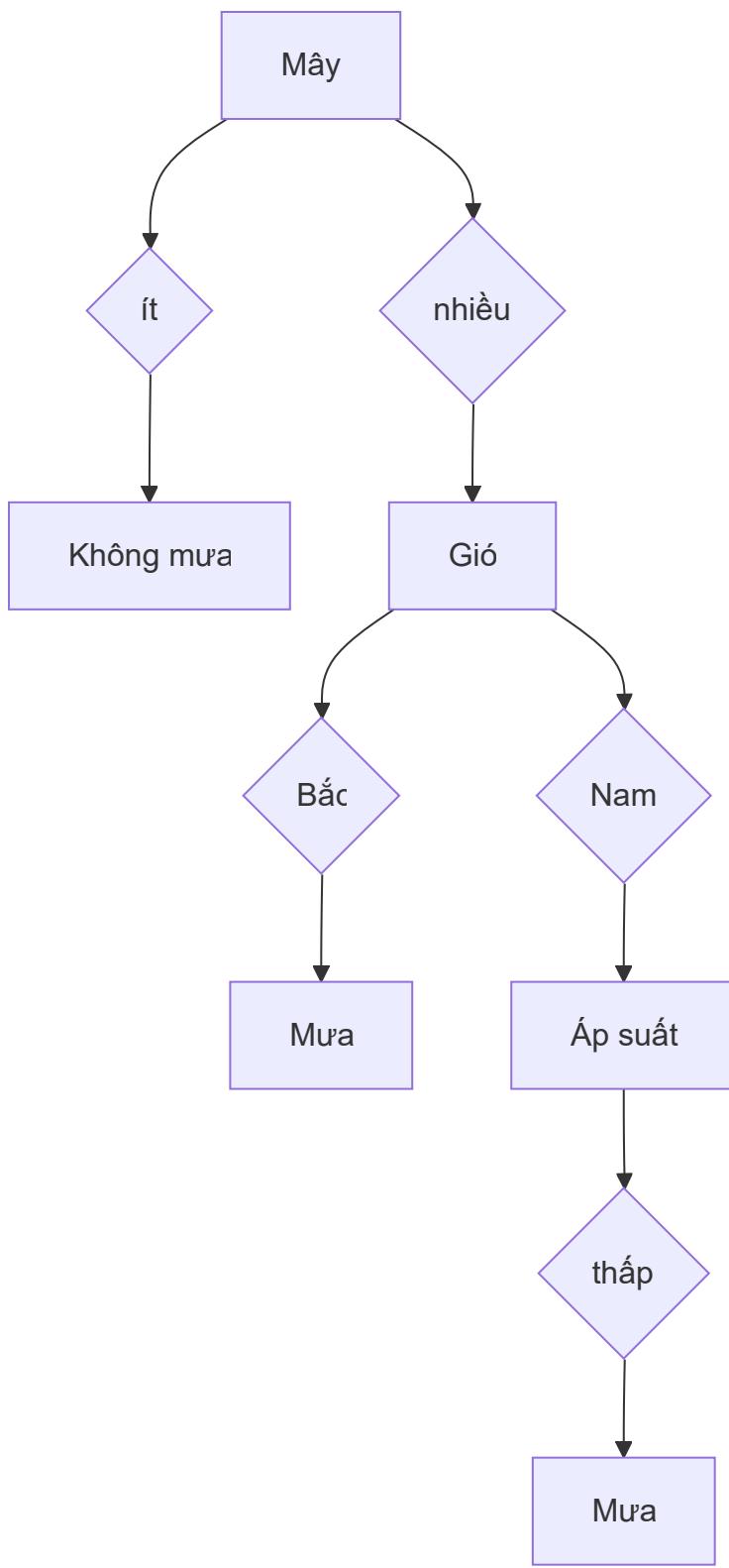
$$Gain(\text{Gió}) \approx 1 - \left(\frac{5}{8} \times 0.9710 + \frac{3}{8} \times 0.9183 \right) \approx 1 - (0.6069 + 0.3444) = 1 - 0.9513 = 0.0$$

2. Chọn thuộc tính gốc và xây dựng cây:

- Thuộc tính 'Mây' có *Gain* cao nhất (0.5488), chọn 'Mây' làm nút gốc.
- **Nhánh Mây = ít:** Gồm {1, 4, 8}. Tất cả đều có kết quả 'không mưa'. Đây là nút lá.
- **Nhánh Mây = nhiều:** Gồm {2, 3, 5, 6, 7}. Có 4 M, 1 KM (*Entropy* ≈ 0.7219). Cần tách tiếp.
 - **Tính Gain cho tập con (Mây=nhiều):** $S' = \{2, 3, 5, 6, 7\}$.
 - **Thuộc tính 'Áp suất' (trong S'):**
 - cao ({6}): 1 M, 0 KM $\rightarrow E = 0$.
 - trung bình ({3}): 1 M, 0 KM $\rightarrow E = 0$.
 - thấp ({2, 5, 7}): 2 M, 1 KM $\rightarrow E \approx 0.9183$.
 - $Gain(\text{Áp suất} | \text{Mây=nhiều}) \approx 0.7219 - (\frac{1}{5} \times 0 + \frac{1}{5} \times 0 + \frac{3}{5} \times 0.9183) \approx 0.7219 - 0.5488 = 0.1731$.
 - **Thuộc tính 'Gió' (trong S'):**
 - Bắc ({3, 5, 6}): 3 M, 0 KM $\rightarrow E = 0$.
 - Nam ({2, 7}): 1 M, 1 KM $\rightarrow E = 1$.
 - $Gain(\text{Gió} | \text{Mây=nhiều}) = 0.7219 - (\frac{3}{5} \times 0 + \frac{2}{5} \times 1) = 0.7219 - 0.4 = 0.3219$.
 - Chọn 'Gió' làm nút tách tiếp theo cho nhánh (Mây=nhiều) vì có *Gain* cao hơn.
 - **Nhánh (Mây=nhiều, Gió=Bắc):** Gồm {3, 5, 6}. Tất cả đều 'mưa'. Đây là nút lá.
 - **Nhánh (Mây=nhiều, Gió=Nam):** Gồm {2, 7}. Có 1 M, 1 KM (*Entropy* = 1). Cần tách tiếp bằng thuộc tính còn lại 'Áp suất'.
 - **Nhánh (Mây=nhiều, Gió=Nam, Áp suất=cao):** Không có mẫu nào trong dữ liệu ban đầu. (Xử lý: Có thể gán lớp phỏ biến của nút cha (Mây=nhiều, Gió=Nam) là M hoặc KM - không xác định rõ ràng, hoặc lớp phỏ biến của nút (Mây=nhiều) là 'mưa'). Giả sử gán lớp phỏ biến của nút cha gần nhất (1M, 1KM), không xác định rõ. Tuy nhiên, trong ví dụ này, không có giá trị 'cao' ở đây.
 - **Nhánh (Mây=nhiều, Gió=Nam, Áp suất=trung bình):** Không có mẫu nào.
 - **Nhánh (Mây=nhiều, Gió=Nam, Áp suất=thấp):** Gồm {2, 7}. 1 M, 1 KM. *Entropy* = 1. Không còn thuộc tính để tách. Gán lớp phỏ biến nhất. Vì số lượng bằng nhau (1 M, 1 KM), có thể chọn tùy ý hoặc dựa trên phân phối tổng thể. Giả sử chọn 'mưa' (lớp phỏ biến hơn trong nút cha Mây=nhiều). \rightarrow Nút lá 'mưa'. Lưu ý: Cách xử lý này có thể khác nhau. Nếu chọn 'không mưa' thì luật sẽ khác.

3. Cây quyết định và luật phân lớp:

- **Cây quyết định (dựa trên các bước trên):**



- Luật phân lớp (giả sử nút lá cuối là 'mưa'):

1. IF Mây = ít THEN Kết quả = không mưa
 2. IF Mây = nhiều AND Gió = Bắc THEN Kết quả = mưa
 3. IF Mây = nhiều AND Gió = Nam AND Áp suất = thấp THEN Kết quả = mưa
- (Lưu ý: Không có luật cho Mây=nhiều, Gió=Nam, Áp suất != thấp từ cây này)

4. Kết luận:

Dựa trên cây quyết định đã xây dựng, các luật phân lớp là:

- Nếu Mây là 'ít', thì Kết quả là 'không mưa'.
- Nếu Mây là 'nhiều' và Gió là 'Bắc', thì Kết quả là 'mưa'.
- Nếu Mây là 'nhiều', Gió là 'Nam', và Áp suất là 'thấp', thì Kết quả là 'mưa' (hoặc 'không mưa' tùy cách xử lý).

Bài tập 7

Cho CSDL sau:

Bảng dữ liệu:

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Nam	mưa
3	ít	thấp	Bắc	không mưa
4	nhiều	trung bình	Bắc	mưa
5	nhiều	thấp	Nam	không mưa
6	nhiều	thấp	Bắc	mưa
7	ít	cao	Nam	không mưa
8	nhiều	cao	Bắc	mưa

Yêu cầu:

Sử dụng thuật toán **cây quyết định** để tìm các luật phân lớp với cột "Kết quả" là thuộc tính phân lớp. Sử dụng bộ luật phân lớp tìm được để xác định lớp cho các đối tượng mới:

Đối tượng	Mây	Áp suất	Gió	Kết quả
9	ít	trung bình	Bắc	?
10	ít	thấp	Nam	?
11	nhiều	trung bình	Nam	?

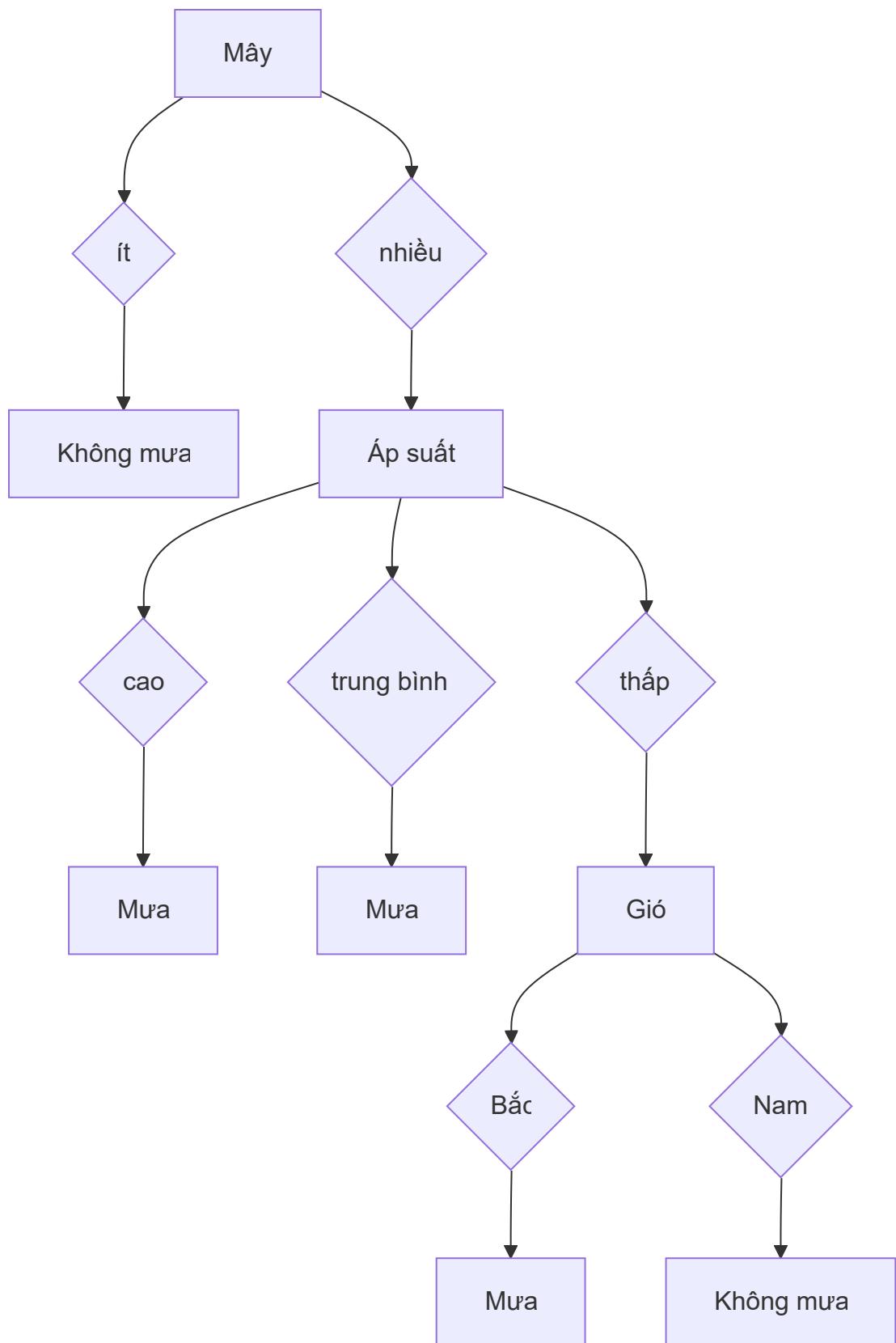
Giải:

1. Xây dựng cây quyết định (tương tự Bài tập 6, sử dụng dữ liệu Bài tập 7):

- **Dữ liệu gốc (S):** 8 đối tượng, 4 'mưa' (M), 4 'không mưa' (KM). $Entropy(S) = 1$.
- **Tính Gain:**
 - $Gain(\text{Mây}) \approx 0.5488$ (Tính toán như ở trên, nhưng dùng dữ liệu Bảng 7)
 - Mây=ít ($\{1, 3, 7\}$): 0 M, 3 KM $\rightarrow E = 0$.
 - Mây=nhiều ($\{2, 4, 5, 6, 8\}$): 4 M, 1 KM $\rightarrow E \approx 0.7219$.
 - $Gain(\text{Áp suất}) \approx 0.2499$ (Tính toán như ở trên, nhưng dùng dữ liệu Bảng 7)
 - cao ($\{1, 2, 8\}$): 2 M, 1 KM $\rightarrow E \approx 0.9183$.
 - trung bình ($\{4\}$): 1 M, 0 KM $\rightarrow E = 0$.
 - thấp ($\{3, 5, 6, 7\}$): 1 M, 3 KM $\rightarrow E \approx 0.8113$.
 - $Gain(\text{Gió}) \approx 0.0487$ (Tính toán như ở trên, nhưng dùng dữ liệu Bảng 7)
 - Bắc ($\{1, 3, 4, 6, 8\}$): 3 M, 2 KM $\rightarrow E \approx 0.9710$.
 - Nam ($\{2, 5, 7\}$): 1 M, 2 KM $\rightarrow E \approx 0.9183$.
- **Chọn nút gốc:** 'Mây' ($Gain$ cao nhất).
- **Nhánh Mây = ít:** Gồm $\{1, 3, 7\}$. Tất cả 'không mưa'. Nút lá **không mưa**.
- **Nhánh Mây = nhiều:** Gồm $\{2, 4, 5, 6, 8\}$. 4 M, 1 KM. $E \approx 0.7219$. Tách tiếp.
 - **Tính Gain cho tập con (Mây=nhiều):** $S' = \{2, 4, 5, 6, 8\}$.
 - **Thuộc tính 'Áp suất' (trong S'):**
 - cao ($\{2, 8\}$): 2 M, 0 KM $\rightarrow E = 0$.
 - trung bình ($\{4\}$): 1 M, 0 KM $\rightarrow E = 0$.
 - thấp ($\{5, 6\}$): 1 M, 1 KM $\rightarrow E = 1$.
 - $Gain(\text{Áp suất} | \text{Mây=nhiều}) = 0.7219 - (\frac{2}{5} \times 0 + \frac{1}{5} \times 0 + \frac{2}{5} \times 1) = 0.7219 - 0.4 = 0.3219$.
 - **Thuộc tính 'Gió' (trong S'):**
 - Bắc ($\{4, 6, 8\}$): 3 M, 0 KM $\rightarrow E = 0$.
 - Nam ($\{2, 5\}$): 1 M, 1 KM $\rightarrow E = 1$.
 - $Gain(\text{Gió} | \text{Mây=nhiều}) = 0.7219 - (\frac{3}{5} \times 0 + \frac{2}{5} \times 1) = 0.7219 - 0.4 = 0.3219$.
- Cả 'Áp suất' và 'Gió' có $Gain$ bằng nhau. Chọn một thuộc tính, ví dụ 'Áp suất'.
- **Nhánh (Mây=nhiều, Áp suất=cao):** Gồm $\{2, 8\}$. Tất cả 'mưa'. Nút lá **mưa**.
- **Nhánh (Mây=nhiều, Áp suất=trung bình):** Gồm $\{4\}$. 'mưa'. Nút lá **mưa**.
- **Nhánh (Mây=nhiều, Áp suất=thấp):** Gồm $\{5, 6\}$. 1 M, 1 KM. $E = 1$. Tách tiếp bằng 'Gió'.
 - **Nhánh (Mây=nhiều, Áp suất=thấp, Gió=Bắc):** Gồm $\{6\}$. 'mưa'. Nút lá **mưa**.
 - **Nhánh (Mây=nhiều, Áp suất=thấp, Gió=Nam):** Gồm $\{5\}$. 'không mưa'. Nút lá **không mưa**.

2. Cây quyết định và luật phân lớp:

- Cây quyết định (chọn Áp suất trước):



- Luật phân lớp:

1. IF Mây = ít THEN Kết quả = không mưa
2. IF Mây = nhiều AND Áp suất = cao THEN Kết quả = mưa
3. IF Mây = nhiều AND Áp suất = trung bình THEN Kết quả = mưa

4. IF Mây = nhiều AND Áp suất = thấp AND Gió = Bắc THEN Kết quả = mưa
5. IF Mây = nhiều AND Áp suất = thấp AND Gió = Nam THEN Kết quả = không mưa

3. Phân lớp đối tượng mới:

- **Đối tượng 9 (Mây=ít, Áp suất=trung bình, Gió=Bắc):**
 - Áp dụng luật 1: Mây = ít -> Kết quả = không mưa.
- **Đối tượng 10 (Mây=ít, Áp suất=thấp, Gió=Nam):**
 - Áp dụng luật 1: Mây = ít -> Kết quả = không mưa.
- **Đối tượng 11 (Mây=nhiều, Áp suất=trung bình, Gió=Nam):**
 - Áp dụng luật 3: Mây = nhiều AND Áp suất = trung bình -> Kết quả = mưa.

4. Kết luận:

Dựa trên bộ luật phân lớp từ cây quyết định:

- Đối tượng 9 được phân lớp là **không mưa**.
- Đối tượng 10 được phân lớp là **không mưa**.
- Đối tượng 11 được phân lớp là **mưa**.

6. Phân lớp Naïve Bayes

Bài tập 8

Cho CSDL sau:

Bảng dữ liệu:

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Bắc	mưa
3	ít	thấp	Bắc	không mưa
4	nhiều	thấp	Bắc	mưa
5	nhiều	trung bình	Bắc	mưa
6	ít	cao	Nam	không mưa
7	nhiều	cao	Nam	mưa
8	nhiều	thấp	Nam	không mưa

Yêu cầu:

Sử dụng thuật toán **Naïve Bayes** để xác định lớp cho mẫu mới sau:

Đối tượng	Mây	Áp suất	Gió	Kết quả
9	ít	thấp	Nam	?
10	ít	trung bình	Bắc	?
11	nhiều	cao	Bắc	?
12	nhiều	trung bình	Nam	?

Giải:

Áp dụng thuật toán Naive Bayes với [Laplace Smoothing](#) ($\alpha = 1$) để tránh xác suất bằng 0.

1. Tính xác suất tiên nghiệm $P(C_k)$:

Tổng số mẫu $N = 8$.

Số lớp 'mưa' = 4. Số lớp 'không mưa' = 4.

$$P(\text{mưa}) = \frac{4}{8} = 0.5$$

$$P(\text{không mưa}) = \frac{4}{8} = 0.5$$

2. Tính xác suất có điều kiện $P(X_i = x_i | C_k)$ sử dụng Laplace:

Công thức: $P(X_i = x_i | C_k) = \frac{\text{count}(X_i=x_i, C_k)+1}{\text{count}(C_k)+V_i}$

Trong đó:

- $V_{\text{Mây}} = 2$ (ít, nhiều)
- $V_{\text{Áp suất}} = 3$ (cao, thấp, trung bình)
- $V_{\text{Gió}} = 2$ (Bắc, Nam)
- **Lớp = mưa (count = 4):**

- $P(\text{Mây}=ít|\text{mưa}) = \frac{0+1}{4+2} = \frac{1}{6}$
- $P(\text{Mây}=nhiều|\text{mưa}) = \frac{4+1}{4+2} = \frac{5}{6}$
- $P(\text{Áp suất}=cao|\text{mưa}) = \frac{2+1}{4+3} = \frac{3}{7}$
- $P(\text{Áp suất}=thấp|\text{mưa}) = \frac{1+1}{4+3} = \frac{2}{7}$
- $P(\text{Áp suất}=trung bình|\text{mưa}) = \frac{1+1}{4+3} = \frac{2}{7}$
- $P(\text{Gió}=Bắc|\text{mưa}) = \frac{3+1}{4+2} = \frac{4}{6} = \frac{2}{3}$
- $P(\text{Gió}=Nam|\text{mưa}) = \frac{1+1}{4+2} = \frac{2}{6} = \frac{1}{3}$

- **Lớp = không mưa (count = 4):**

- $P(\text{Mây}=ít|\text{không mưa}) = \frac{3+1}{4+2} = \frac{4}{6} = \frac{2}{3}$
- $P(\text{Mây}=nhiều|\text{không mưa}) = \frac{1+1}{4+2} = \frac{2}{6} = \frac{1}{3}$
- $P(\text{Áp suất}=cao|\text{không mưa}) = \frac{2+1}{4+3} = \frac{3}{7}$
- $P(\text{Áp suất}=thấp|\text{không mưa}) = \frac{2+1}{4+3} = \frac{3}{7}$

- $P(\text{Áp suất}=trung bình|không mưa) = \frac{0+1}{4+3} = \frac{1}{7}$
- $P(\text{Gió}=Bắc|không mưa) = \frac{2+1}{4+2} = \frac{3}{6} = \frac{1}{2}$
- $P(\text{Gió}=Nam|không mưa) = \frac{2+1}{4+2} = \frac{3}{6} = \frac{1}{2}$

3. Phân lớp các mẫu mới:

- **Mẫu 9: X = (Mây=ít, Áp suất=thấp, Gió=Nam)**
 - $\text{Score}(\text{mưa}) \propto P(\text{mưa}) \times P(\text{Mây}=ít|\text{mưa}) \times P(\text{Áp suất}=thấp|\text{mưa}) \times P(\text{Gió}=Nam|\text{mưa})$
 $\propto 0.5 \times \frac{1}{6} \times \frac{2}{7} \times \frac{1}{3} = \frac{1}{126} \approx 0.0079$
 - $\text{Score}(\text{không mưa}) \propto P(\text{không mưa}) \times P(\text{Mây}=ít|\text{không mưa}) \times P(\text{Áp suất}=thấp|\text{không mưa})$
 $\propto 0.5 \times \frac{2}{3} \times \frac{3}{7} \times \frac{1}{2} = \frac{3}{42} = \frac{1}{14} \approx 0.0714$
 - **Kết luận:** $\text{Score}(\text{không mưa}) > \text{Score}(\text{mưa}) \Rightarrow \text{Mẫu 9: không mưa}$
- **Mẫu 10: X = (Mây=ít, Áp suất=trung bình, Gió=Bắc)**
 - $\text{Score}(\text{mưa}) \propto P(\text{mưa}) \times P(\text{Mây}=ít|\text{mưa}) \times P(\text{Áp suất}=trung bình|\text{mưa}) \times P(\text{Gió}=Bắc|\text{mưa})$
 $\propto 0.5 \times \frac{1}{6} \times \frac{2}{7} \times \frac{2}{3} = \frac{2}{126} = \frac{1}{63} \approx 0.0159$
 - $\text{Score}(\text{không mưa}) \propto P(\text{không mưa}) \times P(\text{Mây}=ít|\text{không mưa}) \times P(\text{Áp suất}=trung bình|\text{không mưa})$
 $\propto 0.5 \times \frac{2}{3} \times \frac{1}{7} \times \frac{1}{2} = \frac{1}{42} \approx 0.0238$
 - **Kết luận:** $\text{Score}(\text{không mưa}) > \text{Score}(\text{mưa}) \Rightarrow \text{Mẫu 10: không mưa}$
- **Mẫu 11: X = (Mây=nhiều, Áp suất=cao, Gió=Bắc)**
 - $\text{Score}(\text{mưa}) \propto P(\text{mưa}) \times P(\text{Mây}=nhiều|\text{mưa}) \times P(\text{Áp suất}=cao|\text{mưa}) \times P(\text{Gió}=Bắc|\text{mưa})$
 $\propto 0.5 \times \frac{5}{6} \times \frac{3}{7} \times \frac{2}{3} = \frac{15}{126} = \frac{5}{42} \approx 0.1190$
 - $\text{Score}(\text{không mưa}) \propto P(\text{không mưa}) \times P(\text{Mây}=nhiều|\text{không mưa}) \times P(\text{Áp suất}=cao|\text{không mưa})$
 $\propto 0.5 \times \frac{1}{3} \times \frac{3}{7} \times \frac{1}{2} = \frac{1.5}{42} = \frac{1}{28} \approx 0.0357$
 - **Kết luận:** $\text{Score}(\text{mưa}) > \text{Score}(\text{không mưa}) \Rightarrow \text{Mẫu 11: mưa}$
- **Mẫu 12: X = (Mây=nhiều, Áp suất=trung bình, Gió=Nam)**
 - $\text{Score}(\text{mưa}) \propto P(\text{mưa}) \times P(\text{Mây}=nhiều|\text{mưa}) \times P(\text{Áp suất}=trung bình|\text{mưa}) \times P(\text{Gió}=Nam|\text{mưa})$
 $\propto 0.5 \times \frac{5}{6} \times \frac{2}{7} \times \frac{1}{3} = \frac{5}{126} \approx 0.0397$
 - $\text{Score}(\text{không mưa}) \propto P(\text{không mưa}) \times P(\text{Mây}=nhiều|\text{không mưa}) \times P(\text{Áp suất}=trung bình|\text{không mưa})$
 $\propto 0.5 \times \frac{1}{3} \times \frac{1}{7} \times \frac{1}{2} = \frac{0.5}{42} = \frac{1}{84} \approx 0.0119$
 - **Kết luận:** $\text{Score}(\text{mưa}) > \text{Score}(\text{không mưa}) \Rightarrow \text{Mẫu 12: mưa}$

7. Phân lớp k-Nearest Neighbors (k-NN)

Bài tập 9

Cho các mẫu dữ liệu được phân bổ trong không gian hai chiều Oxy như hình vẽ 1. Ví dụ: điểm P1 ở tọa độ (3,12). Giả sử người ta tiến hành gán nhãn cho mỗi điểm như sau:

- p1(3, 12): xanh
- p2(9, 13): xanh

- p3(10, 11): đỏ
- p4(9, 8): xanh
- p5(6, 10): đỏ
- p6(3, 9): xanh
- p7(6, 6): đỏ
- p8(5, 4): đỏ
- p9(8, 4): xanh

(Dữ liệu tọa độ được suy ra từ mô tả và hình ảnh)

Hình 1: Phân bố các điểm dữ liệu trong không gian Oxy (Mô tả lại dựa trên hình ảnh)
(Hình ảnh cho thấy sự phân bố của các điểm p1 đến p9, và vị trí của p10(6, 12), p11(8, 6))

Yêu cầu:

Sử dụng thuật toán **k-NN** với khoảng cách **Euclidean** để phân lớp 2 mẫu sau: p10(6, 12) , p11(8, 6) với số lân cận $k = 3$. Thể hiện việc tính toán đầy đủ.

Gợi ý: Công thức Euclidean của 2 điểm A(x_A, y_A), B(x_B, y_B) trong không gian Oxy:

$$AB = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

Giải:

Thuật toán k-NN (k-Nearest Neighbors) phân lớp một điểm dữ liệu mới dựa trên lớp của k điểm láng giềng gần nhất trong tập dữ liệu huấn luyện. Ta sẽ sử dụng $k=3$ và khoảng cách Euclidean.

1. Phân lớp điểm p10(6, 12)

- **Bước 1:** Tính khoảng cách Euclidean từ p10 đến tất cả các điểm huấn luyện p1 đến p9.
 - $d(p10, p1(3, 12)) = \sqrt{(6 - 3)^2 + (12 - 12)^2} = \sqrt{3^2 + 0^2} = \sqrt{9} = 3.0$ (Xanh)
 - $d(p10, p2(9, 13)) = \sqrt{(6 - 9)^2 + (12 - 13)^2} = \sqrt{(-3)^2 + (-1)^2} = \sqrt{9 + 1} = \sqrt{10} \approx 3.16$ (Xanh)
 - $d(p10, p3(10, 11)) = \sqrt{(6 - 10)^2 + (12 - 11)^2} = \sqrt{(-4)^2 + 1^2} = \sqrt{16 + 1} = \sqrt{17} \approx 4.12$ (Đỏ)
 - $d(p10, p4(9, 8)) = \sqrt{(6 - 9)^2 + (12 - 8)^2} = \sqrt{(-3)^2 + 4^2} = \sqrt{9 + 16} = \sqrt{25} = 5.0$ (Xanh)
 - $d(p10, p5(6, 10)) = \sqrt{(6 - 6)^2 + (12 - 10)^2} = \sqrt{0^2 + 2^2} = \sqrt{4} = 2.0$ (Đỏ)
 - $d(p10, p6(3, 9)) = \sqrt{(6 - 3)^2 + (12 - 9)^2} = \sqrt{3^2 + 3^2} = \sqrt{9 + 9} = \sqrt{18} \approx 4.24$ (Xanh)
 - $d(p10, p7(6, 6)) = \sqrt{(6 - 6)^2 + (12 - 6)^2} = \sqrt{0^2 + 6^2} = \sqrt{36} = 6.0$ (Đỏ)
 - $d(p10, p8(5, 4)) = \sqrt{(6 - 5)^2 + (12 - 4)^2} = \sqrt{1^2 + 8^2} = \sqrt{1 + 64} = \sqrt{65} \approx 8.06$ (Đỏ)

- $d(p10, p9(8, 4)) = \sqrt{(6 - 8)^2 + (12 - 4)^2} = \sqrt{(-2)^2 + 8^2} = \sqrt{4 + 64} = \sqrt{68} \approx 8.25$ (Xanh)
- Bước 2:** Xác định 3 láng giềng gần nhất (k=3).
Sắp xếp các điểm theo khoảng cách tăng dần:

1. p5 (Đỏ, $d = 2.0$)
2. p1 (Xanh, $d = 3.0$)
3. p2 (Xanh, $d \approx 3.16$)
4. p3 (Đỏ, $d \approx 4.12$)
5. ...

3 láng giềng gần nhất là: p5 (Đỏ), p1 (Xanh), p2 (Xanh).

- Bước 3:** Xác định lớp đa số trong 3 láng giềng.
Có 2 điểm lớp 'Xanh' và 1 điểm lớp 'Đỏ'. Lớp đa số là 'Xanh'.
- Kết luận:** Điểm $p10(6, 12)$ được phân lớp là **Xanh**.

2. Phân lớp điểm $p11(8, 6)$

- Bước 1:** Tính khoảng cách Euclidean từ $p11$ đến tất cả các điểm huấn luyện $p1$ đến $p9$.
 - $d(p11, p1(3, 12)) = \sqrt{(8 - 3)^2 + (6 - 12)^2} = \sqrt{5^2 + (-6)^2} = \sqrt{25 + 36} = \sqrt{61} \approx 7.81$ (Xanh)
 - $d(p11, p2(9, 13)) = \sqrt{(8 - 9)^2 + (6 - 13)^2} = \sqrt{(-1)^2 + (-7)^2} = \sqrt{1 + 49} = \sqrt{50} \approx 7.07$ (Xanh)
 - $d(p11, p3(10, 11)) = \sqrt{(8 - 10)^2 + (6 - 11)^2} = \sqrt{(-2)^2 + (-5)^2} = \sqrt{4 + 25} = \sqrt{29} \approx 5.39$ (Đỏ)
 - $d(p11, p4(9, 8)) = \sqrt{(8 - 9)^2 + (6 - 8)^2} = \sqrt{(-1)^2 + (-2)^2} = \sqrt{1 + 4} = \sqrt{5} \approx 2.24$ (Xanh)
 - $d(p11, p5(6, 10)) = \sqrt{(8 - 6)^2 + (6 - 10)^2} = \sqrt{2^2 + (-4)^2} = \sqrt{4 + 16} = \sqrt{20} \approx 4.47$ (Đỏ)
 - $d(p11, p6(3, 9)) = \sqrt{(8 - 3)^2 + (6 - 9)^2} = \sqrt{5^2 + (-3)^2} = \sqrt{25 + 9} = \sqrt{34} \approx 5.83$ (Xanh)
 - $d(p11, p7(6, 6)) = \sqrt{(8 - 6)^2 + (6 - 6)^2} = \sqrt{2^2 + 0^2} = \sqrt{4} = 2.0$ (Đỏ)
 - $d(p11, p8(5, 4)) = \sqrt{(8 - 5)^2 + (6 - 4)^2} = \sqrt{3^2 + 2^2} = \sqrt{9 + 4} = \sqrt{13} \approx 3.61$ (Đỏ)
 - $d(p11, p9(8, 4)) = \sqrt{(8 - 8)^2 + (6 - 4)^2} = \sqrt{0^2 + 2^2} = \sqrt{4} = 2.0$ (Xanh)
- Bước 2:** Xác định 3 láng giềng gần nhất (k=3).
Sắp xếp các điểm theo khoảng cách tăng dần:

1. p7 (Đỏ, $d = 2.0$)
2. p9 (Xanh, $d = 2.0$)
3. p4 (Xanh, $d \approx 2.24$)
4. p8 (Đỏ, $d \approx 3.61$)
5. ...

3 láng giềng gần nhất là: $p7$ (Đỏ), $p9$ (Xanh), $p4$ (Xanh).

- **Bước 3:** Xác định lớp đa số trong 3 láng giềng.

Có 2 điểm lớp 'Xanh' và 1 điểm lớp 'Đỏ'. Lớp đa số là 'Xanh'.

- **Kết luận:** Điểm $p11(8, 6)$ được phân lớp là **Xanh**.

8. Phân lớp dựa trên Luật (Rule-Based Classification - Tập trung vào ILA)

Bài tập 10

Sử dụng CSDL từ **Bài tập 7**.

Yêu cầu:

Sử dụng thuật toán **ILA** (Giả định là một thuật toán phân lớp dựa trên luật hoặc ví dụ) để tìm các luật phân lớp với cột "Kết quả" là thuộc tính phân lớp. Sử dụng bộ luật phân lớp tìm được để xác định lớp cho các đối tượng mới:

Đối tượng	Mây	Áp suất	Gió	Kết quả
9	ít	trung bình	Bắc	?
10	ít	thấp	Nam	?
11	nhiều	trung bình	Nam	?

(Lưu ý: So sánh kết quả với câu 7 - thuật toán Cây Quyết Định)

Giải:

Sử dụng thuật toán sinh luật tuần tự (Sequential Covering) như một cách triển khai cho ILA. Mục tiêu là tạo ra một danh sách các luật IF-THEN để phân loại các đối tượng.

1. Học luật cho lớp 'không mưa' (KM):

- **Tập dữ liệu ban đầu:** $D = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Các mẫu 'không mưa': $D_{KM} = \{1, 3, 5, 7\}$.
- **Tìm luật 1:**
 - Xét các điều kiện đơn giản nhất để phủ các mẫu KM:
 - IF Mây = ít THEN KM : Phủ $\{1, 3, 7\}$ (3 KM, 0 M). Độ chính xác = $3/3 = 100\%$.
 - IF Áp suất = thấp AND Gió = Nam THEN KM : Phủ $\{5\}$ (1 KM, 0 M). Độ chính xác = $1/1 = 100\%$.

- Chọn luật có độ phủ lớn nhất và chính xác cao nhất:
Luật 1: IF Mây = ít THEN Kết quả = không mưa
 - Luật này phủ các mẫu {1, 3, 7}. Loại bỏ chúng khỏi D .
- **Tập dữ liệu còn lại:** $D' = \{2, 4, 5, 6, 8\}$. Mẫu KM còn lại: $D'_{KM} = \{5\}$.
- **Tìm luật 2 (để phủ mẫu {5}):**
 - Mẫu {5} có thuộc tính (Mây=nhiều, Áp suất=thấp, Gió=Nam).
 - Xét các điều kiện phủ {5}:
 - IF Mây = nhiều AND Áp suất = thấp AND Gió = Nam THEN KM : Phù {5} (1 KM, 0 M). Độ chính xác = 100%.
 - IF Áp suất = thấp AND Gió = Nam THEN KM : Phù {5} (1 KM, 0 M). Độ chính xác = 100%. (Đơn giản hơn).
 - Chọn luật đơn giản và chính xác:
- **Luật 2: IF Áp suất = thấp AND Gió = Nam THEN Kết quả = không mưa**
- Luật này phủ mẫu {5}. Loại bỏ nó khỏi D' .
- **Tập dữ liệu còn lại:** $D'' = \{2, 4, 6, 8\}$. Không còn mẫu KM nào.

2. Luật mặc định:

- Tất cả các mẫu còn lại trong D'' đều thuộc lớp 'mưa'.
- Thêm luật mặc định:
Luật 3: IF TRUE THEN Kết quả = mưa

3. Bộ luật phân lớp cuối cùng (sắp xếp theo thứ tự ưu tiên):

1. IF Mây = ít THEN Kết quả = không mưa
2. IF Áp suất = thấp AND Gió = Nam THEN Kết quả = không mưa
3. IF TRUE THEN Kết quả = mưa

4. Phân lớp đối tượng mới:

- **Đối tượng 9 (Mây=ít, Áp suất=trung bình, Gió=Bắc):**
 - Thỏa mãn Luật 1 (Mây = ít).
 - **Kết quả = không mưa.**
- **Đối tượng 10 (Mây=ít, Áp suất=thấp, Gió=Nam):**
 - Thỏa mãn Luật 1 (Mây = ít).
 - **Kết quả = không mưa.**
- **Đối tượng 11 (Mây=nhiều, Áp suất=trung bình, Gió=Nam):**
 - Không thỏa mãn Luật 1 ($Mây \neq$ ít).
 - Không thỏa mãn Luật 2 ($\text{Áp suất} \neq$ thấp hoặc $\text{Gió} \neq$ Nam).
 - Áp dụng Luật 3 (Mặc định).

- Kết quả = mưa.

5. So sánh với kết quả của Cây Quyết Định (Bài tập 7):

Kết quả phân lớp cho các đối tượng mới bằng thuật toán ILA (Sequential Covering) là:

- Đối tượng 9: **không mưa**
- Đối tượng 10: **không mưa**
- Đối tượng 11: **mưa**

Kết quả này **giống hệt** với kết quả thu được từ thuật toán Cây Quyết Định trong giải pháp của Bài tập 7.

9. Đánh giá Mô hình Phân lớp (Model Evaluation)

Bài tập 11

Cho bảng dữ liệu thống kê kết quả của một thuật toán phân lớp số khách hàng đến siêu thị có mua hay không mua sản phẩm trong 1 tháng:

Thông kê kết quả:

	Lớp dự đoán	
Lớp thực sự	Mua	Không mua
Mua	8986	1009
Không mua	1358	2547

Yêu cầu:

- Lập ma trận sai số (confusion matrix).
- Tính các độ đo: accuracy, error rate, sensitivity, specificity, precision.

Giải:

1. Lập Ma trận Nhầm Lẫn (Confusion Matrix):

Dựa vào bảng thống kê, ta xác định các giá trị True Positive (TP), False Negative (FN), False Positive (FP), và True Negative (TN) với lớp "Mua" là lớp Positive và "Không mua" là lớp Negative:

- **TP (Mua thực tế, dự đoán Mua):** $TP = 8986$

- **FN (Mua thực tế, dự đoán Không mua):** $FN = 1009$
- **FP (Không mua thực tế, dự đoán Mua):** $FP = 1358$
- **TN (Không mua thực tế, dự đoán Không mua):** $TN = 2547$

Ma trận nhầm lẫn:

	Dự đoán: Mua	Dự đoán: Không mua	Tổng thực tế
Thực tế: Mua	$TP = 8986$	$FN = 1009$	9995
Thực tế: Không mua	$FP = 1358$	$TN = 2547$	3905
Tổng dự đoán	10344	3556	N=13900

2. Tính các độ đo:

- **Tổng số mẫu:** $N = TP + FN + FP + TN = 8986 + 1009 + 1358 + 2547 = 13900$
- **Độ Chính Xác (Accuracy):**

$$\text{Accuracy} = \frac{TP + TN}{N} = \frac{8986 + 2547}{13900} = \frac{11533}{13900} \approx 0.8297 \text{ (hay } 82.97\text{)}$$

- **Tỷ Lệ Lỗi (Error Rate):**

$$\text{Error Rate} = \frac{FP + FN}{N} = \frac{1358 + 1009}{13900} = \frac{2367}{13900} \approx 0.1703 \text{ (hay } 17.03\text{)}$$

(Hoặc: $1 - \text{Accuracy} = 1 - 0.8297 = 0.1703$)

- **Độ Nhạy (Sensitivity / Recall / True Positive Rate - TPR):** (Tỷ lệ khách hàng mua thực tế được dự đoán đúng)

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{8986}{8986 + 1009} = \frac{8986}{9995} \approx 0.8990 \text{ (hay } 89.90\text{)}$$

- **Độ Đặc Hiệu (Specificity / True Negative Rate - TNR):** (Tỷ lệ khách hàng không mua thực tế được dự đoán đúng)

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{2547}{2547 + 1358} = \frac{2547}{3905} \approx 0.6522 \text{ (hay } 65.22\text{)}$$

- **Độ Chuẩn Xác (Precision / Positive Predictive Value - PPV):** (Tỷ lệ khách hàng được dự đoán là mua thực sự mua)

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{8986}{8986 + 1358} = \frac{8986}{10344} \approx 0.8687 \text{ (hay } 86.87\text{)}$$

Kết quả:

- Ma trận nhầm lẫn đã được lập như trên.

- Accuracy $\approx 82.97\%$
- Error Rate $\approx 17.03\%$
- Sensitivity (Recall) $\approx 89.90\%$
- Specificity $\approx 65.22\%$
- Precision $\approx 86.87\%$

Phần 4: Phân cụm (Clustering)

10. Phân cụm Phân cấp (Hierarchical Clustering - AGNES)

Bài tập 12

Cho tập dữ liệu gồm 7 điểm trong không gian 2 chiều: P1, P2, P3, P4, P5, P6, P7.

Cho ma trận khoảng cách giữa các điểm như trong bảng 1.

Bảng 1. Ma trận khoảng cách:

	P1	P2	P3	P4	P5	P6	P7
P1	0.00	0.27	0.23	0.56	0.17	0.40	0.14
P2	0.27	0.00	0.06	0.75	0.33	0.25	0.26
P3	0.23	0.06	0.00	0.59	0.28	0.24	0.22
P4	0.56	0.75	0.59	0.00	0.44	0.48	0.46
P5	0.17	0.33	0.28	0.44	0.00	0.37	0.09
P6	0.40	0.25	0.24	0.48	0.37	0.00	0.31
P7	0.14	0.26	0.22	0.46	0.09	0.31	0.00

Yêu cầu:

a) Sử dụng thuật toán AGNES

Hãy sử dụng lần lượt thuật toán AGNES với **Single link** và **Complete link** để gom nhóm (trình bày chi tiết các bước). Vẽ sơ đồ hình cây (dendrogram) cho kết quả gom nhóm. (Sơ đồ hình cây phải vẽ rõ ràng để nhận biết được thứ tự và giá trị của vị trí các NHÓM gộp lại với nhau.)

b) Xác định nhóm và so sánh

Dựa trên sơ đồ hình cây tương ứng (dùng Single Link/ Complete Link) xác định **3 nhóm** thu được. So sánh kết quả.

Giải:

a) Sử dụng thuật toán AGNES

1. AGNES với Single Link

- **Bước 0:** Ban đầu có 7 cụm: $\{P1\}$, $\{P2\}$, $\{P3\}$, $\{P4\}$, $\{P5\}$, $\{P6\}$, $\{P7\}$. Ma trận khoảng cách là Bảng 1.
- **Bước 1:** Tìm khoảng cách nhỏ nhất trong ma trận: $d(P2, P3) = 0.06$. Gộp $\{P2, P3\}$ thành cụm $C1 = \{P2, P3\}$. Khoảng cách gộp: 0.06.
Các cụm hiện tại: $\{P1\}$, $C1=\{P2, P3\}$, $\{P4\}$, $\{P5\}$, $\{P6\}$, $\{P7\}$.
- **Bước 2:** Cập nhật ma trận khoảng cách (Single Link):
 - $d(C1, P1) = \min(d(P2, P1), d(P3, P1)) = \min(0.27, 0.23) = 0.23$
 - $d(C1, P4) = \min(d(P2, P4), d(P3, P4)) = \min(0.75, 0.59) = 0.59$
 - $d(C1, P5) = \min(d(P2, P5), d(P3, P5)) = \min(0.33, 0.28) = 0.28$
 - $d(C1, P6) = \min(d(P2, P6), d(P3, P6)) = \min(0.25, 0.24) = 0.24$
 - $d(C1, P7) = \min(d(P2, P7), d(P3, P7)) = \min(0.26, 0.22) = 0.22$Ma trận mới:

	P1	C1	P4	P5	P6	P7
P1	0.00	0.23	0.56	0.17	0.40	0.14
C1	0.23	0.00	0.59	0.28	0.24	0.22
P4	0.56	0.59	0.00	0.44	0.48	0.46
P5	0.17	0.28	0.44	0.00	0.37	0.09
P6	0.40	0.24	0.48	0.37	0.00	0.31
P7	0.14	0.22	0.46	0.09	0.31	0.00

- **Bước 3:** Tìm khoảng cách nhỏ nhất: $d(P5, P7) = 0.09$. Gộp $\{P5, P7\}$ thành cụm $C2 = \{P5, P7\}$. Khoảng cách gộp: 0.09.
Các cụm hiện tại: $\{P1\}$, $C1=\{P2, P3\}$, $\{P4\}$, $C2=\{P5, P7\}$, $\{P6\}$.
- **Bước 4:** Cập nhật ma trận khoảng cách (Single Link):
 - $d(C2, P1) = \min(d(P5, P1), d(P7, P1)) = \min(0.17, 0.14) = 0.14$
 - $d(C2, C1) = \min(d(P5, C1), d(P7, C1)) = \min(0.28, 0.22) = 0.22$
 - $d(C2, P4) = \min(d(P5, P4), d(P7, P4)) = \min(0.44, 0.46) = 0.44$
 - $d(C2, P6) = \min(d(P5, P6), d(P7, P6)) = \min(0.37, 0.31) = 0.31$Ma trận mới:

	P1	C1	P4	C2	P6
P1	0.00	0.23	0.56	0.14	0.40
C1	0.23	0.00	0.59	0.22	0.24
P4	0.56	0.59	0.00	0.44	0.48
C2	0.14	0.22	0.44	0.00	0.31
P6	0.40	0.24	0.48	0.31	0.00

- Bước 5:** Tìm khoảng cách nhỏ nhất: $d(C2, P1) = 0.14$. Gộp $\{P1, C2\}$ thành cụm $C3 = \{P1, P5, P7\}$. Khoảng cách gộp: 0.14.

Các cụm hiện tại: $C1=\{P2, P3\}, \{P4\}, C3=\{P1, P5, P7\}, \{P6\}$.

- Bước 6:** Cập nhật ma trận khoảng cách (Single Link):

- $d(C3, C1) = \min(d(P1, C1), d(C2, C1)) = \min(0.23, 0.22) = 0.22$
- $d(C3, P4) = \min(d(P1, P4), d(C2, P4)) = \min(0.56, 0.44) = 0.44$
- $d(C3, P6) = \min(d(P1, P6), d(C2, P6)) = \min(0.40, 0.31) = 0.31$

Ma trận mới:

	C1	P4	C3	P6
C1	0.00	0.59	0.22	0.24
P4	0.59	0.00	0.44	0.48
C3	0.22	0.44	0.00	0.31
P6	0.24	0.48	0.31	0.00

- Bước 7:** Tìm khoảng cách nhỏ nhất: $d(C3, C1) = 0.22$. Gộp $\{C1, C3\}$ thành cụm $C4 = \{P1, P2, P3, P5, P7\}$. Khoảng cách gộp: 0.22.

Các cụm hiện tại: $\{P4\}, C4=\{P1, P2, P3, P5, P7\}, \{P6\}$.

- Bước 8:** Cập nhật ma trận khoảng cách (Single Link):

- $d(C4, P4) = \min(d(C1, P4), d(C3, P4)) = \min(0.59, 0.44) = 0.44$
- $d(C4, P6) = \min(d(C1, P6), d(C3, P6)) = \min(0.24, 0.31) = 0.24$

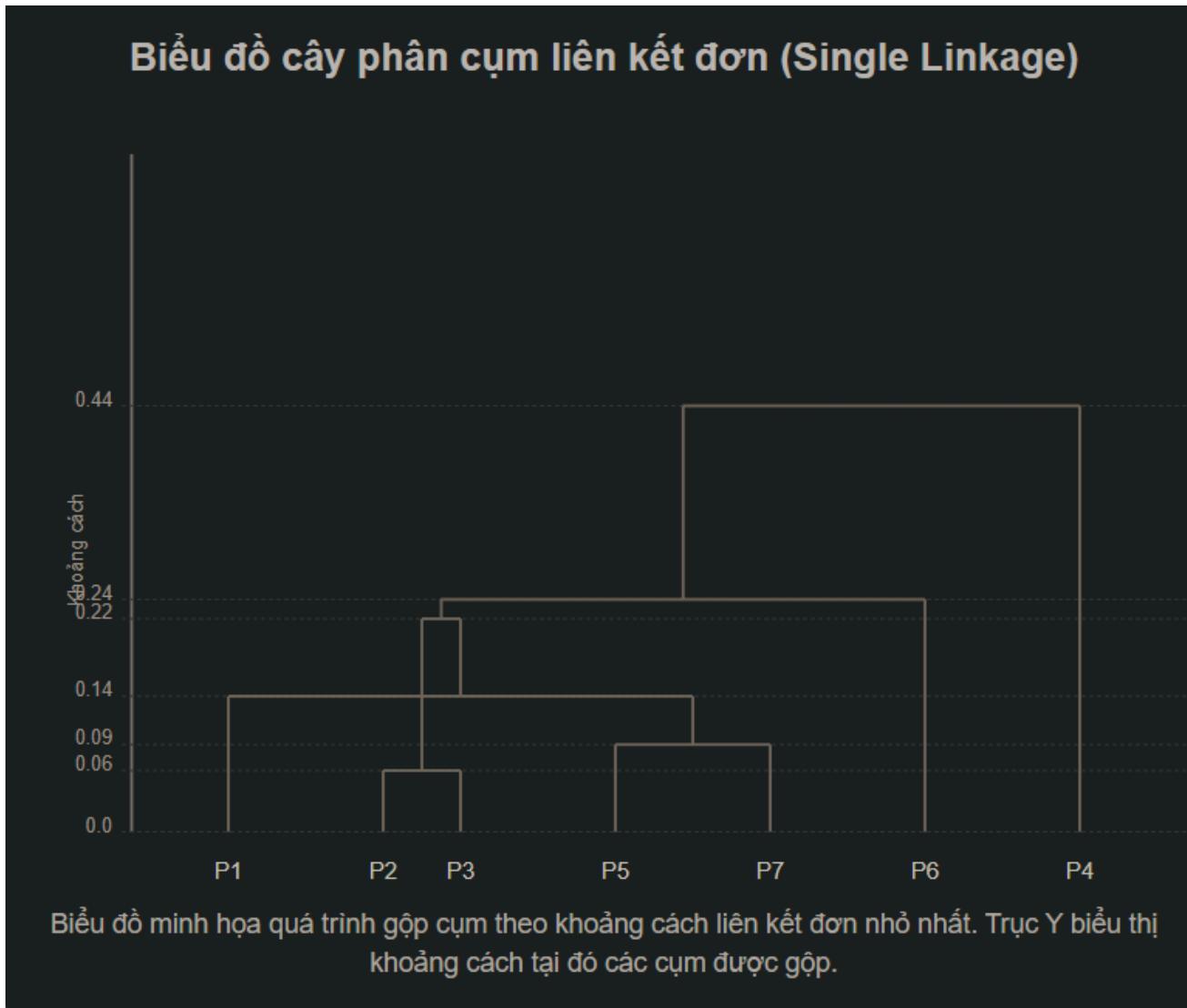
Ma trận mới:

	P4	C4	P6
P4	0.00	0.44	0.48
C4	0.44	0.00	0.24
P6	0.48	0.24	0.00

- **Bước 9:** Tìm khoảng cách nhỏ nhất: $d(C4, P6) = 0.24$. Gộp $\{C4, P6\}$ thành cụm $C5 = \{P1, P2, P3, P5, P6, P7\}$. Khoảng cách gộp: 0.24.
Các cụm hiện tại: $\{P4\}$, $C5 = \{P1, P2, P3, P5, P6, P7\}$.
- **Bước 10:** Cập nhật ma trận khoảng cách (Single Link):
 - $d(C5, P4) = \min(d(C4, P4), d(P6, P4)) = \min(0.44, 0.48) = 0.44$
 - Ma trận mới:

	P4	C5
P4	0.00	0.44
C5	0.44	0.00

- **Bước 11:** Gộp $\{P4, C5\}$ thành cụm $C6 = \{P1, P2, P3, P4, P5, P6, P7\}$. Khoảng cách gộp: 0.44.
- **Dendrogram (Single Linkage):**



Gộp {P5, P7} tại 0.09

Gộp {P1, {P5, P7}} tại 0.14

Gộp {{P2, P3}, {P1, P5, P7}} tại 0.22

Gộp {P6, {P1, P2, P3, P5, P7}} tại 0.24

Gộp {P4, {P1, P2, P3, P5, P6, P7}} tại 0.44

2. AGNES với Complete Link

- **Bước 0:** Ban đầu có 7 cụm: {P1}, {P2}, {P3}, {P4}, {P5}, {P6}, {P7}.
- **Bước 1:** Tìm khoảng cách nhỏ nhất: $d(P2, P3) = 0.06$. Gộp {P2, P3} thành C1 = {P2, P3}. Khoảng cách gộp: 0.06.
Các cụm hiện tại: {P1}, C1={P2, P3}, {P4}, {P5}, {P6}, {P7}.
- **Bước 2:** Cập nhật ma trận khoảng cách (Complete Link):
 - $d(C1, P1) = \max(d(P2, P1), d(P3, P1)) = \max(0.27, 0.23) = 0.27$
 - $d(C1, P4) = \max(d(P2, P4), d(P3, P4)) = \max(0.75, 0.59) = 0.75$
 - $d(C1, P5) = \max(d(P2, P5), d(P3, P5)) = \max(0.33, 0.28) = 0.33$
 - $d(C1, P6) = \max(d(P2, P6), d(P3, P6)) = \max(0.25, 0.24) = 0.25$
 - $d(C1, P7) = \max(d(P2, P7), d(P3, P7)) = \max(0.26, 0.22) = 0.26$Ma trận mới:

	P1	C1	P4	P5	P6	P7
P1	0.00	0.27	0.56	0.17	0.40	0.14
C1	0.27	0.00	0.75	0.33	0.25	0.26
P4	0.56	0.75	0.00	0.44	0.48	0.46
P5	0.17	0.33	0.44	0.00	0.37	0.09
P6	0.40	0.25	0.48	0.37	0.00	0.31
P7	0.14	0.26	0.46	0.09	0.31	0.00

- **Bước 3:** Tìm khoảng cách nhỏ nhất: $d(P5, P7) = 0.09$. Gộp {P5, P7} thành C2 = {P5, P7}. Khoảng cách gộp: 0.09.
Các cụm hiện tại: {P1}, C1={P2, P3}, {P4}, C2={P5, P7}, {P6}.
- **Bước 4:** Cập nhật ma trận khoảng cách (Complete Link):
 - $d(C2, P1) = \max(d(P5, P1), d(P7, P1)) = \max(0.17, 0.14) = 0.17$
 - $d(C2, C1) = \max(d(P5, C1), d(P7, C1)) = \max(0.33, 0.26) = 0.33$
 - $d(C2, P4) = \max(d(P5, P4), d(P7, P4)) = \max(0.44, 0.46) = 0.46$
 - $d(C2, P6) = \max(d(P5, P6), d(P7, P6)) = \max(0.37, 0.31) = 0.37$Ma trận mới:

	P1	C1	P4	C2	P6
P1	0.00	0.27	0.56	0.17	0.40
C1	0.27	0.00	0.75	0.33	0.25
P4	0.56	0.75	0.00	0.46	0.48
C2	0.17	0.33	0.46	0.00	0.37
P6	0.40	0.25	0.48	0.37	0.00

- Bước 5:** Tìm khoảng cách nhỏ nhất: $d(C2, P1) = 0.17$. Gộp $\{P1, C2\}$ thành $C3 = \{P1, P5, P7\}$. Khoảng cách gộp: 0.17.

Các cụm hiện tại: $C1=\{P2, P3\}$, $\{P4\}$, $C3=\{P1, P5, P7\}$, $\{P6\}$.

- Bước 6:** Cập nhật ma trận khoảng cách (Complete Link):

- $d(C3, C1) = \max(d(P1, C1), d(C2, C1)) = \max(0.27, 0.33) = 0.33$
- $d(C3, P4) = \max(d(P1, P4), d(C2, P4)) = \max(0.56, 0.46) = 0.56$
- $d(C3, P6) = \max(d(P1, P6), d(C2, P6)) = \max(0.40, 0.37) = 0.40$

Ma trận mới:

	C1	P4	C3	P6
C1	0.00	0.75	0.33	0.25
P4	0.75	0.00	0.56	0.48
C3	0.33	0.56	0.00	0.40
P6	0.25	0.48	0.40	0.00

- Bước 7:** Tìm khoảng cách nhỏ nhất: $d(C1, P6) = 0.25$. Gộp $\{C1, P6\}$ thành $C4 = \{P2, P3, P6\}$. Khoảng cách gộp: 0.25.

Các cụm hiện tại: $\{P4\}$, $C3=\{P1, P5, P7\}$, $C4=\{P2, P3, P6\}$.

- Bước 8:** Cập nhật ma trận khoảng cách (Complete Link):

- $d(C4, P4) = \max(d(C1, P4), d(P6, P4)) = \max(0.75, 0.48) = 0.75$
- $d(C4, C3) = \max(d(C1, C3), d(P6, C3)) = \max(0.33, 0.40) = 0.40$ (Vì $d(P6, C3) = d(C3, P6) = 0.40$)

Ma trận mới:

	P4	C3	C4
P4	0.00	0.56	0.75
C3	0.56	0.00	0.40

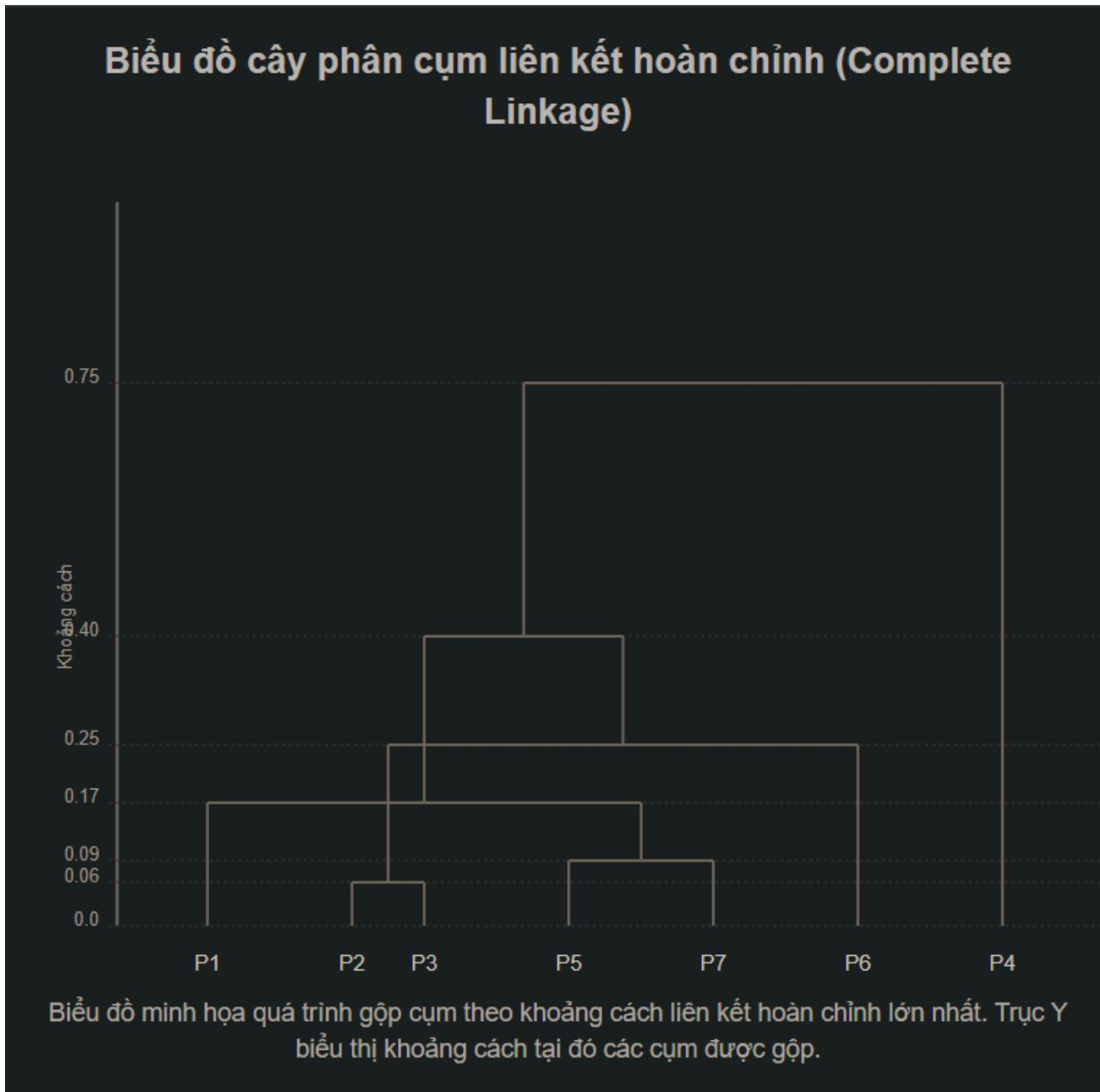
	P4	C3	C4
C4	0.75	0.40	0.00

- **Bước 9:** Tìm khoảng cách nhỏ nhất: $d(C4, C3) = 0.40$. Gộp $\{C3, C4\}$ thành $C5 = \{P1, P2, P3, P5, P6, P7\}$. Khoảng cách gộp: 0.40.
Các cụm hiện tại: $\{P4\}$, $C5=\{P1, P2, P3, P5, P6, P7\}$.
- **Bước 10:** Cập nhật ma trận khoảng cách (Complete Link):
 - $d(C5, P4) = \max(d(C3, P4), d(C4, P4)) = \max(0.56, 0.75) = 0.75$
 Ma trận mới:

	P4	C5
P4	0.00	0.75
C5	0.75	0.00

- **Bước 11:** Gộp $\{P4, C5\}$ thành $C6 = \{P1, P2, P3, P4, P5, P6, P7\}$. Khoảng cách gộp: 0.75.

- Dendrogram (Complete Link):



(Lưu ý: Dendrogram vẽ dạng text có thể không chính xác hoàn toàn về tỷ lệ)

Gộp $\{P_2, P_3\}$ tại 0.06

Gộp $\{P_5, P_7\}$ tại 0.09

Gộp $\{P_1, \{P_5, P_7\}\}$ tại 0.17

Gộp $\{P_6, \{P_2, P_3\}\}$ tại 0.25

Gộp $\{\{P_1, P_5, P_7\}, \{P_2, P_3, P_6\}\}$ tại 0.40

Gộp $\{P_4, \{P_1, P_2, P_3, P_5, P_6, P_7\}\}$ tại 0.75

b) Xác định nhóm và so sánh

- Xác định 3 nhóm (Single Link):

Để có 3 nhóm, chúng ta cần cắt dendrogram Single Link ở mức khoảng cách sao cho có 3 nhánh chính. Cắt ở mức ngay trên 0.22 (ví dụ: 0.23).

- Nhóm 1: {P4}
- Nhóm 2: {P6}
- Nhóm 3: {P1, P2, P3, P5, P7}

- **Xác định 3 nhóm (Complete Link):**

Để có 3 nhóm, chúng ta cắt dendrogram Complete Link ở mức khoảng cách sao cho có 3 nhánh chính. Cắt ở mức ngay trên 0.25 (ví dụ: 0.30).

- Nhóm 1: {P4}
- Nhóm 2: {P1, P5, P7}
- Nhóm 3: {P2, P3, P6}

- **So sánh kết quả:**

- Cả hai phương pháp đều tách điểm P4 thành một cụm riêng biệt.
- **Single Link:** Có xu hướng tạo ra một cụm lớn (<{P1, P2, P3, P5, P7}) và các cụm nhỏ (thậm chí là điểm đơn lẻ như P6). Điều này là do đặc tính "nối cầu" của Single Link, dễ bị ảnh hưởng bởi nhiều hoặc các điểm nằm giữa các cụm.
- **Complete Link:** Tạo ra các cụm có vẻ "chặt chẽ" hơn và kích thước cân bằng hơn (<{P1, P5, P7} và {P2, P3, P6}). Complete Link yêu cầu tất cả các điểm trong hai cụm phải "gần nhau" (theo nghĩa khoảng cách xa nhất), nên có xu hướng tạo ra các cụm hình cầu và ít bị ảnh hưởng bởi nhiều hơn Single Link.

Kết quả phân cụm thành 3 nhóm khác nhau đáng kể giữa hai phương pháp, phản ánh sự khác biệt trong cách đo khoảng cách giữa các cụm.

11. Phân cụm Mật độ (Density-Based Clustering - DBSCAN)

Bài tập 13

Cho tập dữ liệu gồm 8 điểm trong không gian 2 chiều:

- A1=(2, 10)
- A2=(2, 5)
- A3=(8, 4)
- A4=(5, 8)
- A5=(7, 5)
- A6=(6, 4)
- A7=(1, 2)
- A8=(4, 9)

Yêu cầu:

Hãy sử dụng thuật toán **DBSCAN** để gom nhóm với $Eps = 2$ và $Minpts = 2$.

Giải:

1. Xác định tham số:

- Bán kính lân cận: $Eps = 2$
- Số điểm tối thiểu trong lân cận: $MinPts = 2$
- Phương pháp đo khoảng cách: Khoảng Cách Euclidean $d(p, q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$

2. Tính ma trận khoảng cách (làm tròn đến 2 chữ số thập phân):

Point	A1(2,10)	A2(2,5)	A3(8,4)	A4(5,8)	A5(7,5)	A6(6,4)	A7(1,2)	A8(4,9)
A1	0	5.00	8.49	3.61	7.07	7.21	8.06	2.24
A2	5.00	0	6.08	4.24	5.00	4.12	3.16	4.47
A3	8.49	6.08	0	5.00	1.41	2.00	7.28	6.40
A4	3.61	4.24	5.00	0	3.61	4.12	7.21	1.41
A5	7.07	5.00	1.41	3.61	0	1.41	6.71	5.00
A6	7.21	4.12	2.00	4.12	1.41	0	5.39	5.39
A7	8.06	3.16	7.28	7.21	6.71	5.39	0	7.62
A8	2.24	4.47	6.40	1.41	5.00	5.39	7.62	0

3. Thực hiện thuật toán DBSCAN:

- Xét A1 (2, 10):
 - Lân cận $N_{Eps}(A1) = \{P | d(A1, P) \leq 2\} = \{A1\}$.
 - $|N_{Eps}(A1)| = 1 < MinPts$. A1 là Điểm nhiễu (Noise) (tạm thời).
- Xét A2 (2, 5):
 - $N_{Eps}(A2) = \{A2\}$.
 - $|N_{Eps}(A2)| = 1 < MinPts$. A2 là Điểm nhiễu (Noise) (tạm thời).
- Xét A3 (8, 4):
 - $N_{Eps}(A3) = \{A3, A5, A6\}$ (vì $d(A3, A5) \approx 1.41 \leq 2$, $d(A3, A6) = 2.00 \leq 2$).
 - $|N_{Eps}(A3)| = 3 \geq MinPts$. A3 là Điểm lõi (Core Point). Tạo **Cụm 1 = {A3}**.
 - Hàng đợi (Seed Set) $S = N_{Eps}(A3) \setminus \{A3\} = \{A5, A6\}$.
- Mở rộng Cụm 1:
 - Lấy A5 từ S. A5 chưa được xét.
 - $N_{Eps}(A5) = \{A3, A5, A6\}$ (vì $d(A5, A3) \approx 1.41 \leq 2$, $d(A5, A6) \approx 1.41 \leq 2$).

- $|N_{Eps}(A5)| = 3 \geq MinPts$. A5 là Điểm lõi.
- Thêm A5 vào Cụm 1. **Cụm 1 = {A3, A5}**.
- Thêm các điểm trong $N_{Eps}(A5)$ chưa xét và chưa có trong S vào S: {A6} đã có. $S = \{A6\}$.
- Lấy **A6** từ S. A6 chưa được xét.
 - $N_{Eps}(A6) = \{A3, A5, A6\}$ (vì $d(A6, A3) = 2.00 \leq 2$, $d(A6, A5) \approx 1.41 \leq 2$).
 - $|N_{Eps}(A6)| = 3 \geq MinPts$. A6 là Điểm lõi.
 - Thêm A6 vào Cụm 1. **Cụm 1 = {A3, A5, A6}**.
 - Thêm các điểm trong $N_{Eps}(A6)$ chưa xét và chưa có trong S vào S: không có. $S = \emptyset$.
- Hàng đợi S rỗng. Kết thúc Cụm 1.
- **Xét A4 (5, 8):**
 - $N_{Eps}(A4) = \{A4, A8\}$ (vì $d(A4, A8) \approx 1.41 \leq 2$).
 - $|N_{Eps}(A4)| = 2 \geq MinPts$. A4 là Điểm lõi. Tạo **Cụm 2 = {A4}**.
 - Hàng đợi S = $N_{Eps}(A4) \setminus \{A4\} = \{A8\}$.
 - **Mở rộng Cụm 2:**
 - Lấy **A8** từ S. A8 chưa được xét.
 - $N_{Eps}(A8) = \{A4, A8\}$ (vì $d(A8, A4) \approx 1.41 \leq 2$).
 - $|N_{Eps}(A8)| = 2 \geq MinPts$. A8 là Điểm lõi.
 - Thêm A8 vào Cụm 2. **Cụm 2 = {A4, A8}**.
 - Thêm các điểm trong $N_{Eps}(A8)$ chưa xét và chưa có trong S vào S: không có. $S = \emptyset$.
 - Hàng đợi S rỗng. Kết thúc Cụm 2.
- **Xét A5 (7, 5):** Đã thuộc Cụm 1. Bỏ qua.
- **Xét A6 (6, 4):** Đã thuộc Cụm 1. Bỏ qua.
- **Xét A7 (1, 2):**
 - $N_{Eps}(A7) = \{A7\}$.
 - $|N_{Eps}(A7)| = 1 < MinPts$. A7 là Điểm nhiễu (Noise). (Kiểm tra: A7 không nằm trong lân cận Eps của bất kỳ Điểm lõi nào ($\{A3, A5, A6, A4, A8\}$) nên không phải là Điểm biên).
- **Xét A8 (4, 9):** Đã thuộc Cụm 2. Bỏ qua.

4. Kết luận:

- **Cụm 1:** {A3, A5, A6} (Các điểm lõi)
- **Cụm 2:** {A4, A8} (Các điểm lõi)
- **Điểm nhiễu (Noise):** {A1, A2, A7}

Okay, chúng ta sẽ cùng nhau giải chi tiết đề thi này.

Đề thi GK 22-23

Câu 1: Phân Tích Dữ Liệu Tuổi và Tỷ Lệ Mỡ

Giả sử một bệnh viện kiểm tra dữ liệu độ tuổi và tỉ lệ mỡ cơ thể của 18 người trưởng thành được chọn ngẫu nhiên với kết quả như sau:

Age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.3	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Tổng số mẫu: $N = 18$.

Dữ liệu:

- **Age:**
- **%fat:** [9.5, 26.3, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2, 34.6, 42.5, 28.8, 33.4, 30.2, 34.1, 32.9, 41.2, 35.7]

Sắp xếp dữ liệu:

- **Sorted Age:**
- **Sorted %fat:** [7.8, 9.5, 17.8, 25.9, 26.3, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5]

a. Tính trung bình, trung vị và độ lệch chuẩn của age và %fat

1. Biến Age:

- **Trung bình (Mean):**

$$\mu_{Age} = \frac{\sum Age_i}{N} = \frac{23 + 23 + \dots + 60 + 61}{18} = \frac{836}{18} \approx 46.44$$

- **Trung vị (Median):** Do $N = 18$ (chẵn), trung vị là trung bình của 2 giá trị ở giữa (thứ 9 và 10) trong dãy đã sắp xếp.

Vị trí: $N/2 = 9$ và $N/2 + 1 = 10$.

Giá trị thứ 9 là 50, giá trị thứ 10 là 52.

$$Median_{Age} = \frac{50 + 52}{2} = 51$$

- **Độ lệch chuẩn (Standard Deviation):**

Phương sai (Variance): $\sigma_{Age}^2 = \frac{\sum (Age_i - \mu_{Age})^2}{N-1}$

$$\sigma_{Age}^2 = \frac{(23 - 46.44)^2 + \dots + (61 - 46.44)^2}{17} \approx \frac{2611.11}{17} \approx 153.59$$

Độ lệch chuẩn:

$$\sigma_{Age} = \sqrt{\sigma_{Age}^2} \approx \sqrt{153.59} \approx 12.39$$

2. Biến %fat:

- **Trung bình (Mean):**

$$\mu_{\%fat} = \frac{\sum \%fat_i}{N} = \frac{9.5 + 26.3 + \dots + 41.2 + 35.7}{18} = \frac{517.9}{18} \approx 28.77$$

- **Trung vị (Median):** Giá trị thứ 9 và 10 trong dãy %fat đã sắp xếp.

Giá trị thứ 9 là 30.2, giá trị thứ 10 là 31.2.

$$Median_{\%fat} = \frac{30.2 + 31.2}{2} = 30.7$$

- **Độ lệch chuẩn (Standard Deviation):**

Phương sai (Variance): $\sigma_{\%fat}^2 = \frac{\sum (\%fat_i - \mu_{\%fat})^2}{N-1}$

$$\sigma_{\%fat}^2 = \frac{(9.5 - 28.77)^2 + \dots + (35.7 - 28.77)^2}{17} \approx \frac{1408.9}{17} \approx 82.88$$

Độ lệch chuẩn:

$$\sigma_{\%fat} = \sqrt{\sigma_{\%fat}^2} \approx \sqrt{82.88} \approx 9.10$$

Kết quả:

- **Age:** Mean ≈ 46.44 , Median = 51, Std Dev ≈ 12.39
- **%fat:** Mean ≈ 28.77 , Median = 30.7, Std Dev ≈ 9.10

b. Vẽ Boxplot của age và %fat

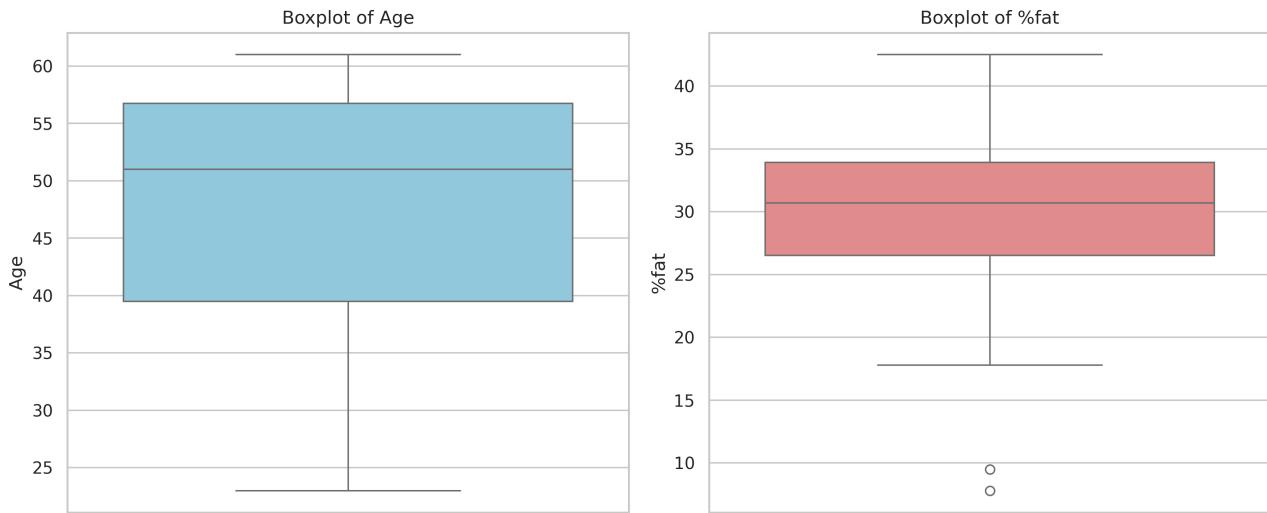
Boxplot cần các giá trị: Minimum, Q1 (Tứ phân vị thứ nhất), Median (Q2), Q3 (Tứ phân vị thứ ba), Maximum. $IQR = Q3 - Q1$. Outlier là các giá trị nằm ngoài khoảng $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$.

1. Boxplot cho Age:

- Sorted Age:
- Median (Q2) = 51
- Q1: Trung vị của nửa dưới. Vị trí giữa là 39. $Q1 = 39$.
- Q3: Trung vị của nửa trên. Vị trí giữa là 57. $Q3 = 57$.
- $IQR = Q3 - Q1 = 57 - 39 = 18$.
- Min không outlier = $Q1 - 1.5 \times IQR = 39 - 1.5 \times 18 = 39 - 27 = 12$. Giá trị nhỏ nhất trong dữ liệu là 23, lớn hơn 12. Vậy Min = 23.
- Max không outlier = $Q3 + 1.5 \times IQR = 57 + 1.5 \times 18 = 57 + 27 = 84$. Giá trị lớn nhất trong dữ liệu là 61, nhỏ hơn 84. Vậy Max = 61.
- Không có outliers.
- **Vẽ Boxplot Age:** Hộp từ 39 đến 57, vạch ở giữa tại 51. "Râu" kéo dài đến 23 và 61.

2. Boxplot cho %fat:

- Sorted %fat: [7.8, 9.5, 17.8, 25.9, 26.3, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5]
- Median (Q2) = 30.7
- Q1: Trung vị của nửa dưới [7.8, 9.5, 17.8, 25.9, 26.3, 27.2, 27.4, 28.8, 30.2]. Vị trí giữa là 26.3. $Q1 = 26.3$.
- Q3: Trung vị của nửa trên [31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5]. Vị trí giữa là 34.1. $Q3 = 34.1$.
- $IQR = Q3 - Q1 = 34.1 - 26.3 = 7.8$.
- Min không outlier = $Q1 - 1.5 \times IQR = 26.3 - 1.5 \times 7.8 = 26.3 - 11.7 = 14.6$. Giá trị nhỏ nhất là 7.8 và 9.5, đều nhỏ hơn 14.6. Đây là outliers. Giá trị nhỏ nhất không phải outlier là 17.8.
- Max không outlier = $Q3 + 1.5 \times IQR = 34.1 + 1.5 \times 7.8 = 34.1 + 11.7 = 45.8$. Giá trị lớn nhất là 42.5, nhỏ hơn 45.8. Vậy Max = 42.5.
- Có 2 outliers: 7.8 và 9.5.
- **Vẽ Boxplot %fat:** Hộp từ 26.3 đến 34.1, vạch ở giữa tại 30.7. "Râu" dưới kéo dài đến 17.8, "Râu" trên kéo dài đến 42.5. Đánh dấu 2 điểm outlier tại 7.8 và 9.5.



c. Vẽ scatter plot và q-q plot dựa trên 2 biến này

- **Scatter Plot (Biểu đồ phân tán):**

- Trục hoành (x): Age
- Trục tung (y): %fat
- Mỗi điểm trên biểu đồ tương ứng với một cặp (Age, %fat) của một người. Ví dụ: (23, 9.5), (23, 26.3), ..., (61, 35.7).
- Biểu đồ này giúp hình dung mối quan hệ tuyến tính (hoặc phi tuyến) giữa Age và %fat. Nhìn sơ bộ, có vẻ có xu hướng tuổi tăng thì %fat cũng tăng.

- **Q-Q Plot (Quantile-Quantile Plot):**

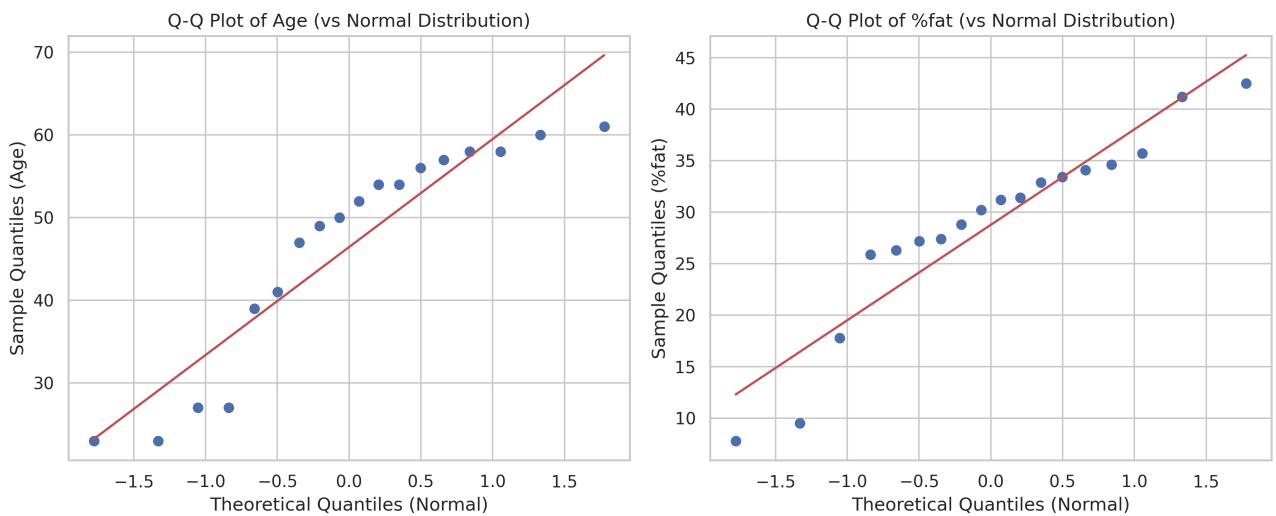
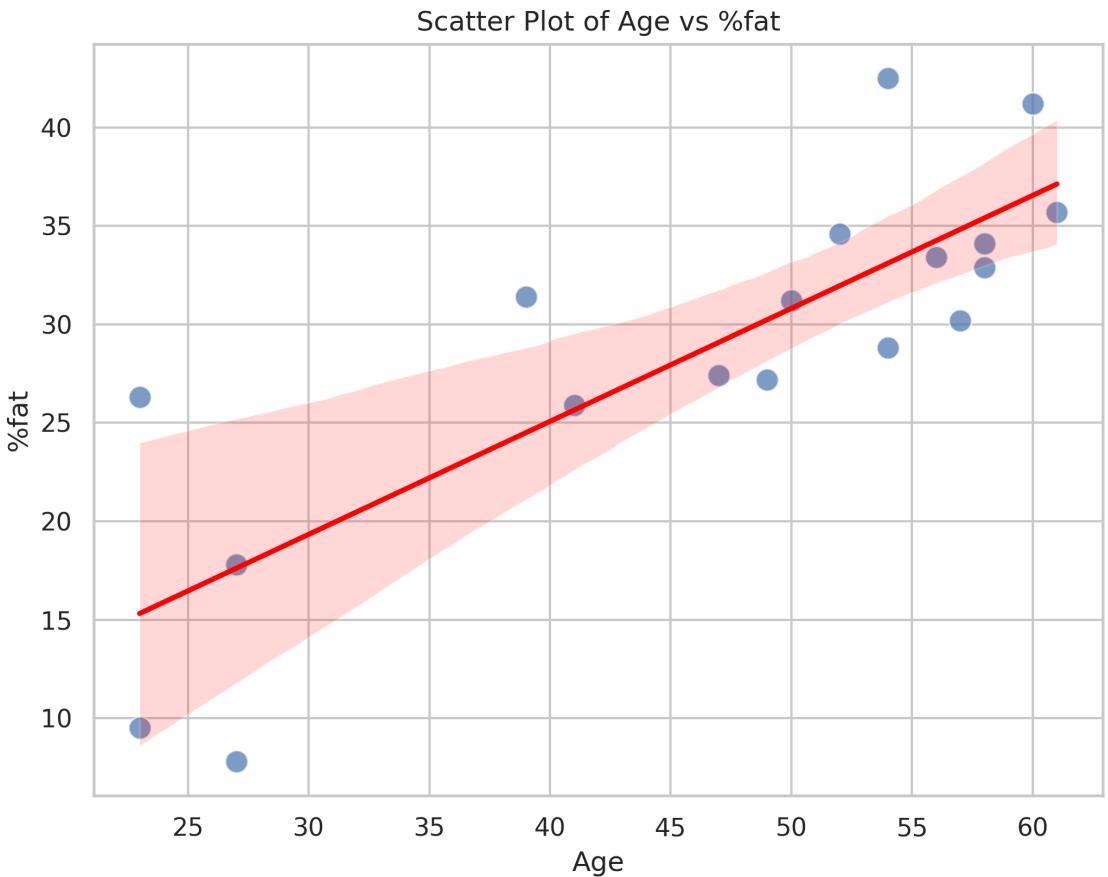
Mục đích: Kiểm tra xem dữ liệu của một biến có tuân theo một phân phối lý thuyết nào đó không (thường là phân phối chuẩn - Normal distribution).

Cách vẽ (cho biến Age):

1. Sắp xếp dữ liệu Age: [23, 23, ..., 61].
2. Tính các lượng tử (quantiles) của dữ liệu Age.
3. Tính các lượng tử tương ứng của phân phối chuẩn lý thuyết (Standard Normal Z-scores).
4. Vẽ biểu đồ với lượng tử lý thuyết trên trục hoành và lượng tử dữ liệu trên trục tung.

Điễn giải: Nếu các điểm nằm gần đường thẳng $y = x$ (hoặc một đường thẳng), dữ liệu có khả năng tuân theo phân phối chuẩn. Nếu các điểm lệch xa đường thẳng, dữ liệu không tuân theo phân phối chuẩn.

Tương tự, vẽ Q-Q plot cho biến %fat.



d. Chuẩn hóa hai biến này dựa trên chuẩn hóa z-score

Công thức chuẩn hóa z-score: $z = \frac{x-\mu}{\sigma}$

Sử dụng các giá trị μ và σ đã tính ở câu a:

- $\mu_{Age} \approx 46.44, \sigma_{Age} \approx 12.39$
- $\mu_{\%fat} \approx 28.77, \sigma_{\%fat} \approx 9.10$

Ví dụ tính z-score cho người đầu tiên (Age=23, %fat=9.5):

- $z_{Age} = \frac{23-46.44}{12.39} \approx -1.89$
- $z_{\%fat} = \frac{9.5-28.77}{9.10} \approx -2.12$

Tính toán cho tất cả các điểm (làm tròn 2 chữ số thập phân):

Age	%fat	z_{Age}	$z_{\%fat}$
23	9.5	-1.89	-2.12
23	26.3	-1.89	-0.27
27	7.8	-1.57	-2.30
27	17.8	-1.57	-1.21
39	31.4	-0.60	0.29
41	25.9	-0.44	-0.32
47	27.4	0.05	-0.15
49	27.2	0.21	-0.17
50	31.2	0.29	0.27
52	34.6	0.45	0.64
54	42.5	0.61	1.51
54	28.8	0.61	0.00
56	33.4	0.77	0.51
57	30.2	0.85	0.16
58	34.1	0.93	0.59
58	32.9	0.93	0.45
60	41.2	1.09	1.37
61	35.7	1.18	0.76

e. Tính hệ số tương quan Pearson. Hai biến này tương quan dương hay âm?

Công thức hệ số tương quan Pearson:

$$r = \frac{\sum_{i=1}^N (Age_i - \mu_{Age})(\%fat_i - \mu_{\%fat})}{\sqrt{\sum_{i=1}^N (Age_i - \mu_{Age})^2 \sum_{i=1}^N (\%fat_i - \mu_{\%fat})^2}}$$

Hoặc sử dụng z-scores:

$$r = \frac{\sum_{i=1}^N z_{Age,i} \times z_{\%fat,i}}{N - 1}$$

Tính tích $z_{Age} \times z_{\%fat}$ cho từng người:

$(-1.89)(-2.12) = 4.01$, $(-1.89)(-0.27) = 0.51$, $(-1.57)(-2.30) = 3.61$, $(-1.57)(-1.21) = 1.90$, $(-0.60)(0.29) = -0.17$, $(-0.44)(-0.32) = 0.14$, $(0.05)(-0.15) = -0.01$, $(0.21)(-0.17) = -0.04$, $(0.29)(0.27) = 0.08$, $(0.45)(0.64) = 0.29$, $(0.61)(1.51) = 0.92$, $(0.61)(0.00) = 0.00$, $(0.77)(0.51) = 0.39$, $(0.85)(0.16) = 0.14$, $(0.93)(0.59) = 0.55$, $(0.93)(0.45) = 0.42$, $(1.09)(1.37) = 1.49$, $(1.18)(0.76) = 0.90$

Tổng các tích: $\sum z_{Age,i} \times z_{\%fat,i} \approx 15.13$

Hệ số tương quan Pearson:

$$r = \frac{15.13}{18 - 1} = \frac{15.13}{17} \approx 0.89$$

Kết luận:

Hệ số tương quan Pearson $r \approx 0.89$.

Vì $r > 0$ và giá trị khá gần 1, hai biến **Age** và **%fat** có **tương quan dương mạnh**. Điều này có nghĩa là khi tuổi tăng lên, tỷ lệ mỡ cơ thể cũng có xu hướng tăng lên trong tập dữ liệu này.

Câu 2: Khai Thác Tập Phổ Biến và Luật Kết Hợp

Cho Cơ sở dữ liệu (CSDL) giao dịch sau và $MinSup = 60\%$.

CSDL Giao dịch:

TID	Items
10	D, H, C, A, B, K, M
20	E, H, D, G, P, I
30	B, C, D, G, H, K
40	E, A, C, B, P, I
50	K, B, M, F, H, D

Xác định ngưỡng hỗ trợ tuyệt đối:

CSDL có 5 giao dịch ($|D| = 5$).

Ngưỡng hỗ trợ tối thiểu tương đối $MinSup = 60\%$.

Ngưỡng hỗ trợ tối thiểu tuyệt đối $minsup_count = 5 \times 60\% = 3$.

a. Áp dụng thuật toán Charm để tìm toàn bộ tập phổ biến đóng trong dữ liệu trên.

Thuật toán CHARM tìm tập phổ biến đóng (Closed Frequent Itemsets - CFI).

Bước 1: Tìm tập phổ biến 1-itemset (L_1) và Tidset của chúng.

Quét CSDL lần 1:

- A: $\{10, 40\}$ -> Sup=2 (< 3) -> Loại
- B: $\{10, 30, 40, 50\}$ -> Sup=4 (≥ 3) -> Giữ lại. Tidset(B) = $\{10, 30, 40, 50\}$
- C: $\{10, 30, 40\}$ -> Sup=3 (≥ 3) -> Giữ lại. Tidset(C) = $\{10, 30, 40\}$
- D: $\{10, 20, 30, 50\}$ -> Sup=4 (≥ 3) -> Giữ lại. Tidset(D) = $\{10, 20, 30, 50\}$
- E: $\{20, 40\}$ -> Sup=2 (< 3) -> Loại
- F: $\{50\}$ -> Sup=1 (< 3) -> Loại
- G: $\{20, 30\}$ -> Sup=2 (< 3) -> Loại
- H: $\{10, 20, 30, 50\}$ -> Sup=4 (≥ 3) -> Giữ lại. Tidset(H) = $\{10, 20, 30, 50\}$
- I: $\{20, 40\}$ -> Sup=2 (< 3) -> Loại
- K: $\{10, 30, 50\}$ -> Sup=3 (≥ 3) -> Giữ lại. Tidset(K) = $\{10, 30, 50\}$
- M: $\{10, 50\}$ -> Sup=2 (< 3) -> Loại
- P: $\{20, 40\}$ -> Sup=2 (< 3) -> Loại

Các 1-itemset phổ biến và tidset của chúng (sắp xếp theo thứ tự từ điển):

- B: $\{10, 30, 40, 50\}$ (Sup=4)
- C: $\{10, 30, 40\}$ (Sup=3)
- D: $\{10, 20, 30, 50\}$ (Sup=4)
- H: $\{10, 20, 30, 50\}$ (Sup=4)
- K: $\{10, 30, 50\}$ (Sup=3)

Bước 2: Khởi tạo các lớp tương đương ban đầu (dựa trên itemset)

Các nút ban đầu cho hàm đệ quy CHARM-Extend :

```
Nodes = [(B, {10,30,40,50}), (C, {10,30,40}), (D, {10,20,30,50}), (H, {10,20,30,50}), (K, {10,30,50})]
```

Bước 3: Gọi đệ quy CHARM-Extend(Nodes)

- Xét cặp (B, C):
 - $X_{ij} = \{B, C\}$
 - $t(X_{ij}) = t(B) \cap t(C) = \{10, 30, 40, 50\} \cap \{10, 30, 40\} = \{10, 30, 40\}$
 - Sup({B,C}) = 3 (≥ 3).
 - Kiểm tra Pruning:

- $t(B) \neq t(C)$
- $t(B) \not\supseteq t(C)$
- $t(C) \subset t(B)$? Không.
- Lưu nút mới (BC, {10, 30, 40}).
- **Xét cặp (B, D):**
 - $X_{ij} = \{B, D\}$
 - $t(X_{ij}) = \{10, 30, 40, 50\} \cap \{10, 20, 30, 50\} = \{10, 30, 50\}$
 - $\text{Sup}(\{B,D\}) = 3 (\geq 3)$.
 - Kiểm tra Pruning: Không áp dụng.
 - Lưu nút mới (BD, {10, 30, 50}).
- **Xét cặp (B, H):**
 - $X_{ij} = \{B, H\}$
 - $t(X_{ij}) = \{10, 30, 40, 50\} \cap \{10, 20, 30, 50\} = \{10, 30, 50\}$
 - $\text{Sup}(\{B,H\}) = 3 (\geq 3)$.
 - Kiểm tra Pruning: Không áp dụng.
 - Lưu nút mới (BH, {10, 30, 50}).
- **Xét cặp (B, K):**
 - $X_{ij} = \{B, K\}$
 - $t(X_{ij}) = \{10, 30, 40, 50\} \cap \{10, 30, 50\} = \{10, 30, 50\}$
 - $\text{Sup}(\{B,K\}) = 3 (\geq 3)$.
 - Kiểm tra Pruning: $t(K) \subset t(B)$? Không.
 - Lưu nút mới (BK, {10, 30, 50}).
- **Xét cặp (C, D):**
 - $X_{ij} = \{C, D\}$
 - $t(X_{ij}) = \{10, 30, 40\} \cap \{10, 20, 30, 50\} = \{10, 30\}$
 - $\text{Sup}(\{C,D\}) = 2 (< 3)$. Loại.
- **Xét cặp (C, H):**
 - $X_{ij} = \{C, H\}$
 - $t(X_{ij}) = \{10, 30, 40\} \cap \{10, 20, 30, 50\} = \{10, 30\}$
 - $\text{Sup}(\{C,H\}) = 2 (< 3)$. Loại.
- **Xét cặp (C, K):**
 - $X_{ij} = \{C, K\}$
 - $t(X_{ij}) = \{10, 30, 40\} \cap \{10, 30, 50\} = \{10, 30\}$
 - $\text{Sup}(\{C,K\}) = 2 (< 3)$. Loại.
- **Xét cặp (D, H):**
 - $X_{ij} = \{D, H\}$
 - $t(X_{ij}) = \{10, 20, 30, 50\} \cap \{10, 20, 30, 50\} = \{10, 20, 30, 50\}$

- $\text{Sup}(\{\text{D,H}\}) = 4 (\geq 3)$.
- Kiểm tra Pruning: $t(D) = t(H)$.
 - **Thuộc tính 1:** Loại bỏ nút tiền nhiệm X_i (D) khỏi cây. Thay thế D bằng DH trong các kết hợp sau.
 - Lưu nút mới (DH, {10, 20, 30, 50}).
- **Xét cặp (D, K):** (Vì D đã bị thay thế bởi DH, ta xét (DH, K))
 - $X_{ij} = \{\text{D, H, K}\}$
 - $t(X_{ij}) = t(DH) \cap t(K) = \{10, 20, 30, 50\} \cap \{10, 30, 50\} = \{10, 30, 50\}$
 - $\text{Sup}(\{\text{DHK}\}) = 3 (\geq 3)$.
 - Kiểm tra Pruning: $t(K) \subset t(DH)$? Không.
 - Lưu nút mới (DHK, {10, 30, 50}).
- **Xét cặp (H, K):** (H vẫn còn trong cây vì nó chưa bị prune ở bước D,H)
 - $X_{ij} = \{\text{H, K}\}$
 - $t(X_{ij}) = \{10, 20, 30, 50\} \cap \{10, 30, 50\} = \{10, 30, 50\}$
 - $\text{Sup}(\{\text{HK}\}) = 3 (\geq 3)$.
 - Kiểm tra Pruning: $t(K) \subset t(H)$? Không.
 - Lưu nút mới (HK, {10, 30, 50}).

Kết thúc vòng lặp đầu tiên. Các nút/lớp tương đương mới được tạo:

- Lớp 1 (từ B): (BC, {10,30,40}), (BD, {10,30,50}), (BH, {10,30,50}), (BK, {10,30,50})
- Lớp 2 (từ C): Không có nút mới phổ biến.
- Lớp 3 (từ D/DH): (DH, {10,20,30,50}), (DHK, {10,30,50})
- Lớp 4 (từ H): (HK, {10,30,50})
- Lớp 5 (từ K): Không có nút mới.

Các nút ban đầu còn lại (chưa bị prune): B, C, H, K.

Tập hợp các nút hiện tại (potential closed itemsets):

B:{4}, C:{3}, H:{4}, K:{3}, BC:{3}, BD:{3}, BH:{3}, BK:{3}, DH:{4}, DHK:{3}, HK:{3}
(Lưu ý: D đã bị loại)

Bước 4: Gọi đệ quy cho các lớp tương đương mới.

- **CHARM-Extend([(BC, {10,30,40}), (BD, {10,30,50}), (BH, {10,30,50}), (BK, {10,30,50})])**
 - Xét (BC, BD): {B,C,D}, t = {10,30}. Sup=2. Loại.
 - Xét (BC, BH): {B,C,H}, t = {10,30}. Sup=2. Loại.
 - Xét (BC, BK): {B,C,K}, t = {10,30}. Sup=2. Loại.
 - Xét (BD, BH): {B,D,H}, t = {10,30,50}. Sup=3.

- **Pruning:** $t(BD) = t(BH)$. Loại BD. Thay BD bằng BDH.
- Lưu (BDH, {10,30,50}).
- Xét (BD, BK): (Thay BD bằng BDH) Xét (BDH, BK): {B,D,H,K}, t={10,30,50}. Sup=3.
 - **Pruning:** $t(BK) \subset t(BDH)$? Không. $t(BDH) = t(BK)$. Loại BDH. Thay BDH bằng BDHK.
 - Lưu (BDHK, {10,30,50}).
- Xét (BH, BK): (BH chưa bị prune). {B,H,K}, t = {10,30,50}. Sup=3.
 - **Pruning:** $t(BH) = t(BK)$. Loại BH. Thay BH bằng BHK.
 - Lưu (BHK, {10,30,50}).

Các nút mới từ lớp này: (BDHK, {10,30,50}), (BHK, {10,30,50}).

Các nút còn lại từ lớp này: BC, BK.

- **CHARM-Extend([(DH, {10,20,30,50}), (DHK, {10,30,50})])**

- Xét (DH, DHK): {D,H,K}, t = {10,30,50}. Sup=3.
 - **Pruning:** $t(DHK) \subset t(DH)$. Thay DHK bằng DHK (không đổi).
 - Lưu (DHK, {10,30,50}).

Các nút mới từ lớp này: (DHK, {10,30,50}).

Các nút còn lại từ lớp này: DH.

- **CHARM-Extend([(HK, {10,30,50})])**: Không có cặp nào để xét.

Tiếp tục gọi đệ quy:

- **CHARM-Extend([(BDHK, {10,30,50}), (BHK, {10,30,50})])**

- Xét (BDHK, BHK): {B,D,H,K}, t={10,30,50}. Sup=3.
 - **Pruning:** $t(BDHK) = t(BHK)$. Loại BDHK. Thay BDHK bằng BDHK.
 - Lưu (BDHK, {10,30,50}).

Các nút mới: (BDHK, {10,30,50}).

Các nút còn lại: BHK.

- **CHARM-Extend([(DHK, {10,30,50})])**: Không có cặp.

Thu thập các nút không bị prune trong quá trình:

Từ các bước trên, các nút được tạo ra và không bị loại bỏ hoàn toàn (hoặc được thay thế và nút thay thế không bị loại) là các tập phỏ biến đóng.

- Bị loại/thay thế: D, BD, BH, BDH, BHK, BDHK (bị thay thế nhưng rồi lại được tạo ra)
- Còn lại/Kết quả cuối cùng của nhánh:
 - Từ gốc: C:{3}, H:{4}, K:{3} (B bị prune bởi các con sau này)

- Từ nhánh B: BC:{3}, BK:{3} (BD, BH bị prune) -> BDHK:{3} (prune BK, BHK)
- Từ nhánh D/DH: DH:{4}, DHK:{3}
- Từ nhánh H: HK:{3}

Kiểm tra lại các nút cuối cùng:

- C: {10, 30, 40} (Sup=3)
- H: {10, 20, 30, 50} (Sup=4) -> Bị prune bởi DH? Không, vì D bị prune trước.
- K: {10, 30, 50} (Sup=3) -> Bị prune bởi BK/DHK/HK?
- BC: {10, 30, 40} (Sup=3)
- BK: {10, 30, 50} (Sup=3) -> Bị prune bởi BDHK?
- DH: {10, 20, 30, 50} (Sup=4)
- DHK: {10, 30, 50} (Sup=3) -> Bị prune bởi BDHK?
- HK: {10, 30, 50} (Sup=3) -> Bị prune bởi BDHK/DHK?
- BDHK: {10, 30, 50} (Sup=3)

Áp dụng lại quy tắc đóng: Một itemset X là đóng nếu không có superset Y nào có cùng support.

- Sup=4: {H}, {D,H}. Chỉ có {D,H} là đóng (vì H là con của DH).
- Sup=3: {C}, {K}, {B,C}, {B,D}, {B,H}, {B,K}, {D,K}, {H,K}, {B,D,H}, {B,D,K}, {B,H,K}, {D,H,K}, {B,D,H,K}.
 - C:{3} vs BC:{3}. C không đóng.
 - K:{3} vs BK:{3}. K không đóng.
 - BC:{3}. Không có cha nào sup=3. BC là đóng.
 - BD:{3} vs BDH:{3}. BD không đóng.
 - BH:{3} vs BDH:{3}. BH không đóng.
 - BK:{3} vs BDK:{3}. BK không đóng.
 - DK:{3} vs DHK:{3}. DK không đóng.
 - HK:{3} vs DHK:{3}. HK không đóng.
 - BDH:{3} vs BDHK:{3}. BDH không đóng.
 - BDK:{3} vs BDHK:{3}. BDK không đóng.
 - BHK:{3} vs BDHK:{3}. BHK không đóng.
 - DHK:{3} vs BDHK:{3}. DHK không đóng.
 - BDHK:{3}. Không có cha nào. BDHK là đóng.

Kết quả cuối cùng - Tập phổ biến đóng (CFI):

- {D, H} (Sup=4)

- {B, C} (Sup=3)
- {B, D, H, K} (Sup=3)

(Kết quả này khớp với kết quả của Apriori trong ví dụ bạn đã cung cấp)

b. Suy ra danh sách toàn bộ tập phỗ biến và độ phỗ biến của mỗi tập.

Từ các tập phỗ biến đóng và support của chúng, ta có thể suy ra các tập phỗ biến con. Một tập con của một tập phỗ biến đóng cũng là phỗ biến. Support của tập con phải lớn hơn hoặc bằng support của tập cha đóng. Tuy nhiên, để có support chính xác, ta cần tính lại hoặc sử dụng kết quả từ thuật toán như Apriori.

Dựa trên kết quả Apriori đã thực hiện trong ví dụ của bạn (vì nó đầy đủ và chính xác):

- L_1 (**Sup** ≥ 3): {B}:4, {C}:3, {D}:4, {H}:4, {K}:3
- L_2 (**Sup** ≥ 3): {B,C}:3, {B,D}:3, {B,H}:3, {B,K}:3, {D,H}:4, {D,K}:3, {H,K}:3
- L_3 (**Sup** ≥ 3): {B,D,H}:3, {B,D,K}:3, {B,H,K}:3, {D,H,K}:3
- L_4 (**Sup** ≥ 3): {B,D,H,K}:3

Danh sách toàn bộ tập phỗ biến và độ phỗ biến:

- **Sup=4**: {B}, {D}, {H}, {D,H}
- **Sup=3**: {C}, {K}, {B,C}, {B,D}, {B,H}, {B,K}, {D,K}, {H,K}, {B,D,H}, {B,D,K}, {B,H,K}, {D,H,K}, {B,D,H,K}

c. Cho biết danh sách luật kết hợp thỏa mãn Minsup=60%?

Để sinh luật kết hợp $X \rightarrow Y$, ta cần $X \cup Y$ là tập phỗ biến và
 $Confidence(X \rightarrow Y) = \frac{Sup(X \cup Y)}{Sup(X)} \geq MinConf.$

Đề bài không cho $MinConf$. Thông thường, nếu không cho, ta có thể liệt kê các luật mạnh (ví dụ: Conf=100%) hoặc các luật có thể sinh ra từ các tập phỗ biến đóng/tối đại.

Sinh luật từ các **tập phỗ biến đóng** tìm được ở câu a:

1. Từ CFI = {D, H} (Sup=4):

- Tập con phỗ biến: {D} (Sup=4), {H} (Sup=4).
- Luật $D \rightarrow H$: $Conf = \frac{Sup(\{D,H\})}{Sup(\{D\})} = \frac{4}{4} = 100\%$
- Luật $H \rightarrow D$: $Conf = \frac{Sup(\{D,H\})}{Sup(\{H\})} = \frac{4}{4} = 100\%$

2. Từ CFI = {B, C} (Sup=3):

- Tập con phỗ biến: {B} (Sup=4), {C} (Sup=3).
- Luật $B \rightarrow C$: $Conf = \frac{Sup(\{B,C\})}{Sup(\{B\})} = \frac{3}{4} = 75\%$
- Luật $C \rightarrow B$: $Conf = \frac{Sup(\{B,C\})}{Sup(\{C\})} = \frac{3}{3} = 100\%$

3. Từ CFI = {B, D, H, K} (Sup=3):

- Các tập con phỗ biến và support của chúng:
 - {B}:4, {D}:4, {H}:4, {K}:3
 - {B,D}:3, {B,H}:3, {B,K}:3, {D,H}:4, {D,K}:3, {H,K}:3
 - {B,D,H}:3, {B,D,K}:3, {B,H,K}:3, {D,H,K}:3
- **Ví dụ một số luật có thể sinh ra:**
 - $\{B, D, H\} \rightarrow K$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{B,D,H\})} = \frac{3}{3} = 100\%$
 - $\{B, D, K\} \rightarrow H$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{B,D,K\})} = \frac{3}{3} = 100\%$
 - $\{B, H, K\} \rightarrow D$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{B,H,K\})} = \frac{3}{3} = 100\%$
 - $\{D, H, K\} \rightarrow B$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{D,H,K\})} = \frac{3}{3} = 100\%$
 - $\{B, D\} \rightarrow \{H, K\}$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{B,D\})} = \frac{3}{3} = 100\%$
 - $\{D, H\} \rightarrow \{B, K\}$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{D,H\})} = \frac{3}{4} = 75\%$
 - $K \rightarrow \{B, D, H\}$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{K\})} = \frac{3}{3} = 100\%$
 - $B \rightarrow \{D, H, K\}$: $Conf = \frac{Sup(\{B,D,H,K\})}{Sup(\{B\})} = \frac{3}{4} = 75\%$
 - ... và nhiều luật khác.

Danh sách các luật kết hợp (ví dụ các luật có Conf $\geq 75\%$):

- $D \rightarrow H$ (Conf=100%)
- $H \rightarrow D$ (Conf=100%)
- $C \rightarrow B$ (Conf=100%)
- $B \rightarrow C$ (Conf=75%)
- $\{B, D, H\} \rightarrow K$ (Conf=100%)
- $\{B, D, K\} \rightarrow H$ (Conf=100%)
- $\{B, H, K\} \rightarrow D$ (Conf=100%)
- $\{D, H, K\} \rightarrow B$ (Conf=100%)
- $\{B, D\} \rightarrow \{H, K\}$ (Conf=100%)
- $\{B, H\} \rightarrow \{D, K\}$ (Conf=100%)
- $\{B, K\} \rightarrow \{D, H\}$ (Conf=100%)
- $\{D, K\} \rightarrow \{B, H\}$ (Conf=100%)
- $\{H, K\} \rightarrow \{B, D\}$ (Conf=100%)
- $K \rightarrow \{B, D, H\}$ (Conf=100%)

- $\{D, H\} \rightarrow \{B, K\}$ (Conf=75%)
- $B \rightarrow \{D, H, K\}$ (Conf=75%)
- $D \rightarrow \{B, H, K\}$ (Conf=75%)
- $H \rightarrow \{B, D, K\}$ (Conf=75%)

(Lưu ý: Việc liệt kê tất cả các luật có thể rất dài. Thông thường sẽ yêu cầu các luật thỏa mãn cả MinSup và MinConf cụ thể)