

โจทย์ข้อที่ 1.1

บริษัทขายไอศกรีมต้องการทำนายยอดขาย (ถ้วย) จากอุณหภูมิสูงสุดของวัน (องศาเซลเซียส) โดยมีข้อมูล

5 วัน

ล่าสุดดังนี้

อุณหภูมิ (X)	ยอดขาย (Y)
25	150
30	200
32	230
28	180
35	250

คำสั่ง:

1. จงหาสมการ Linear Regression ($y=mx+c$) จากข้อมูลข้างต้น

$$\sum xy = (3 \cdot 1.5) + (5 \cdot 2) + (2 \cdot 1) + (6 \cdot 3) + (4 \cdot 2.2) + (7 \cdot 3.5) = 67.8$$

$$\sum x = 25 + 30 + 32 + 28 + 35 = 150$$

$$\sum x^2 = 25^2 + 30^2 + 32^2 + 28^2 + 35^2 = 4558$$

$$\bar{x} = \frac{25 + 30 + 32 + 28 + 35}{5} = 30$$

$$\sum y = 150 + 200 + 230 + 180 + 250 = 1010$$

$$\bar{y} = \frac{(150 + 200 + 230 + 180 + 250)}{5} = 202$$

$$m = \frac{5(30,900) - (150 \cdot 1010)}{5(4558) - (150)^2} = \frac{300}{29} = 10.3448$$

$$c = 202 - (10.3448 \cdot 30) = -108.3448$$

Ans: $y = 10.3448 \cdot x - 108.3448$

2. ถ้าวันนี้อุณหภูมิ 33 องศาเซลเซียส คาดว่าจะขายไอศกรีมได้กี่ถ้วย?

$$10.3448 * 33 - 108.3448 = 233.0342$$

Ans: คาดว่าจะขายไอศกรีมได้ 233.0342 ถ้วย

โจทย์ข้อที่ 1.2

ฟิตเนสแห่งหนึ่งต้องการวิเคราะห์ความสัมพันธ์ระหว่างจำนวนชั่วโมงที่ลูกค้าออกกำลังกายต่อสัปดาห์ (X) กับน้ำหนักที่ลดลงในหนึ่งเดือน (กก.) (Y)

ชั่วโมง/สัปดาห์ (X)	น้ำหนักที่ลด (Y)
3	1.5
5	2.0
2	1.0
6	3.0
4	2.2
7	3.5

คำสั่ง:

1. จงหาสมการ Linear Regression

$$\sum xy = (3 * 1.5) + (5 * 2) + (2 * 1) + (6 * 3) + (4 * 2.2) + (7 * 3.5) = 67.8$$

$$\sum x = 3 + 5 + 2 + 6 + 4 + 7 = 27$$

$$\sum x^2 = 3^2 + 5^2 + 2^2 + 6^2 + 4^2 + 7^2 = 139$$

$$\bar{x} = \frac{3 + 5 + 2 + 6 + 4 + 7}{6} = 4.5$$

$$\sum y = 1.5 + 2 + 1 + 3 + 2.2 + 3.5 = 13.2$$

$$\bar{y} = \frac{(1.5 + 2 + 1 + 3 + 2.2 + 3.5)}{6} = 2.2$$

$$m = \frac{6(67.8) - (27 * 13.2)}{6(139) - (27)^2} = \frac{50.4}{105} = 0.48$$

$$c = 2.2 - (0.48 * 4.5) = 0.04$$

Ans: $y = 0.48 * x + 0.04$

2. หากลูกค้าออกกำลังกาย 8 ชั่วโมง/สัปดาห์ คาดว่าน้ำหนักจะลดลงกี่กิโลกรัม?

$$0.48 * 8 + 0.04 = 3.88$$

Ans: คาดว่าน้ำหนักจะลดลง 4.24 กิโลกรัม

โจทย์ข้อที่ 2.1

ต้องการสร้างโมเดลทำนาย "ราคามือสอง" (Y, หน่วยเป็นพันบาท) ของสมาร์ทโฟน โดยพิจารณาจาก "อายุการใช้งาน (เดือน)" (X1)

อายุ (X1)	ราคา (Y)
6	18
12	14
24	9
8	17
18	11

คำสั่ง: จงหาการแบ่งครั้งแรก (First Split) ที่ดีที่สุด โดยคำนวณค่า Standard Deviation Reduction (SDR) ของทุกจุดแบ่งที่เป็นไปได้

$$\sum y = 18 + 14 + 9 + 17 + 11 = 69$$

$$\bar{y} = \frac{18 + 14 + 9 + 17 + 11}{5} = 13.8$$

$$SD = 1712.409$$

$$SD = 3.4293$$

Unique X

$$(6+8)/2=7$$

$$(8+12)/2=10$$

$$(12+18)/2=15$$

$$(18+24)/2=21$$

$$X \leq 7$$

$$L(y): \{18\}$$

$$\omega_L = 1, \overline{y}_L = 0, SD_L = 0$$

$$R(y): \{14, 9, 17, 11\}$$

$$\omega_R = 4, \overline{y}_R = 3.571, SD_R = 3.031$$

$$SDR = 3.429 - \left[\left(\frac{1}{5} * 0 \right) + \left(\frac{4}{5} * 3.031 \right) \right] = 1.004$$

$$X \leq 10$$

$$L(y): \{18, 17\}$$

$$\omega_L = 2, \overline{y}_L = 17.5, SD_L = 0.5$$

$$R(y): \{14, 9, 11\}$$

$$\omega_R = 3, \overline{y}_R = 11.333, SD_R = 2.055$$

$$SDR = 3.429 - \left[\left(\frac{2}{5} * 0.5 \right) + \left(\frac{3}{5} * 2.055 \right) \right] = 1.996$$

$$X \leq 15$$

$$L(y): \{18, 17, 14\}$$

$$\omega_L = 3, \overline{y}_L = 16.333, SD_L = 1.7$$

$$R(y): \{9, 11\}$$

$$\omega_R = 2, \overline{y}_R = 10, SD_R = 1$$

$$SDR = 3.429 - \left[\left(\frac{3}{5} * 1.7 \right) + \left(\frac{2}{5} * 1 \right) \right] = 2.009$$

$$X \leq 21$$

$$L(y): \{18, 17, 14, 11\}$$

$$\omega_L = 4, \overline{y}_L = 15, SD_L = 2.739$$

$$R(y): \{9\}$$

$$\omega_R = 1, SD_R = 0$$

$$SDR = 3.429 - \left[\left(\frac{4}{5} * 2.739 \right) + \left(\frac{1}{5} * 0 \right) \right] = 1.238$$

Ans: ที่ดีที่สุดคือ $X \leq 15$

$$L(y): \{18, 17, 14\}$$

$$R(y): \{9, 11\}$$

$$\omega_L = 3, \bar{y}_L = 16.333, SD_L = 1.7$$

$$\omega_R = 2, \bar{y}_R = 10, SD_R = 1$$

$$SDR = 3.429 - \left[\left(\frac{3}{5} * 1.7 \right) + \left(\frac{2}{5} * 1 \right) \right] = 2.009$$

โจทย์ข้อที่ 2.2 (โจทย์ท้าทาย)

บริษัทเกมต้องการสร้างโมเดลทำนาย "คะแนนในเกม" (Y) ของผู้เล่น โดยอ้างอิงจาก "ชั่วโมงที่เล่น" (X1) และ "เลเวลผู้เล่น" (X2) เงื่อนไข: หยุดแบ่ง Node (สร้าง Leaf) ก็ต่อเมื่อ Node นั้นมีข้อมูลน้อยกว่าหรือเท่ากับ 3 ชิ้น

ชั่วโมงที่เล่น (X1)	เลเวลผู้เล่น (X2)	คะแนนในเกม (Y)
5	10	1200
15	25	3500
20	30	4500
2	5	500
8	15	1800
25	40	6000
12	20	2800
18	35	4000

1. จงสร้าง Decision Tree จากข้อมูลทั้งหมดให้สมบูรณ์ตามขั้นตอน (แสดงการคำนวณเพื่อหาจุดแบ่งที่ดีที่สุดในแต่ละ Node)

$$\sum y \frac{1200 + 3500 + 4500 + 500 + 1800 + 6000 + 2800 + 4000}{8} = 3037.5$$

$$SD = \sqrt{\frac{(1200 - 3037.5)^2 + (3500 - 3037.5)^2 + (4500 - 3037.5)^2 + (500 - 3037.5)^2 + (1800 - 3037.5)^2 + (6000 - 3037.5)^2 + (2800 - 3037.5)^2 + (4000 - 3037.5)^2}{8}}$$

$$SD = 1712.409$$

UniqueX1

$$(2+5)/2=3.5$$

$$(5+8)/2=6.5$$

$$(8+12)/2=10$$

$$(12+15)/2=13.5$$

$$(15+18)/2=16.5$$

$$(18+20)/2=19$$

$$(20+25)/2=22.5$$

หาจุดแบ่งแรกที่ดีที่สุด

$$X1 \leq 3.5$$

L(y):{500}

$$\omega_L = 1, \overline{y}_L = 0, SD_L = 0$$

UniqueX2

$$(5+10)/2=7.5$$

$$(10+15)/2=12.5$$

$$(15+20)/2=17.5$$

$$(20+25)/2=22.5$$

$$(25+30)/2=27.5$$

$$(30+35)/2=32.5$$

$$(35+40)/2=37.5$$

R(y):{1200,3500,4500,1800,6000,2800,4000}

$$\omega_R = 7, \overline{y}_R = 3400, SD_R = 1516.575$$

$$SDR = 1712.409 - \left[\left(\frac{1}{8} * 0 \right) + \left(\frac{7}{8} * 1516.575 \right) \right] = 385.406$$

$$X1 \leq 6.5$$

L(y):{500,1200}

$$\omega_L = 2, \overline{y}_L = 850, SD_L = 350$$

R(y):{3500,4500,1800,6000,2800,4000}

$$\omega_R = 6, \overline{y}_R = 3766.667, SD_R = 1319.933$$

$$SDR = 1712.409 - \left[\left(\frac{2}{8} * 843.727 \right) + \left(\frac{6}{8} * 1319.933 \right) \right] = 634.959$$

$$X1 \leq 10$$

L(y):{500,1200,1800}

$$\omega_L = 3, \overline{y}_L = 1166.667, SD_L = 531.2459$$

R(y):{3500,4500,6000,2800,4000}

$$\omega_R = 5, \overline{y}_R = 4160, SD_R = 1078.147$$

$$SDR = 1712.409 - \left[\left(\frac{3}{8} * 531.246 \right) + \left(\frac{5}{8} * 1078.147 \right) \right] = 839.35$$

$$X1 \leq 13.5$$

L(y):{500,1200,1800,2800}

$$\omega_L = 4, \overline{y}_L = 1575, SD_L = 843.727$$

R(y):{3500,4500,6000,4000}

$$\omega_R = 4, \overline{y}_R = 4500, SD_R = 935.414$$

$$SDR = 1712.409 - \left[\left(\frac{4}{8} * 843.727 \right) + \left(\frac{4}{8} * 935.414 \right) \right] = 822.823$$

X1<=16.5

L(y):{500,1200,1800,2800,3500}

R(y):{4500, 6000 ,4000}

$\omega_L = 5, \overline{y_L} = 1960, SD_L = 1069.626$

$\omega_R = 3, \overline{y_R} = 4833.333, SD_R = 849.837$

$$SDR = 1712.409 - \left[\left(\frac{5}{8} * 1069.626 \right) + \left(\frac{3}{8} * 849.837 \right) \right] = 725.204$$

X1<=19

L(y):{500,1200,1800,2800,3500,4000}

R(y):{4500, 6000}

$\omega_L = 6, \overline{y_L} = 2300, SD_L = 1226.784$

$\omega_R = 2, \overline{y_R} = 5250, SD_R = 750$

$$SDR = 1712.409 - \left[\left(\frac{6}{8} * 1226.784 \right) + \left(\frac{2}{8} * 750 \right) \right] = 604.821$$

X1<=22.5

L(y):{500,1200,1800,2800,3500,4500}

R(y):{6000}

$\omega_L = 7, \overline{y_L} = 2614.286, SD_L = 1385.051$

$\omega_R = 1, SD_R = 0$

$$SDR = 1712.409 - \left[\left(\frac{7}{8} * 1385.051 \right) + \left(\frac{1}{8} * 0 \right) \right] = 500.489$$

X2<=3.5

L(y):{500}

R(y):{1200,3500,4500,1800,6000,2800,4000}

$\omega_L = 1, SD_L = 0$

$\omega_R = 7, \overline{y_R} = 3400, SD_R = 1516.575$

$$SDR = 1712.409 - \left[\left(\frac{1}{8} * 0 \right) + \left(\frac{7}{8} * 1516.575 \right) \right] = 385.406$$

X2<=12.5

L(y):{500,1200}

R(y):{3500,4500,1800,6000,2800,4000}

$\omega_L = 2, \overline{y_L} = 850, SD_L = 350$

$\omega_R = 6, \overline{y_R} = 3766.667, SD_R = 1319.933$

$$SDR = 1712.409 - \left[\left(\frac{2}{8} * 843.727 \right) + \left(\frac{6}{8} * 1319.933 \right) \right] = 634.959$$

X2<=17.5

L(y):{500,1200,1800}

R(y):{3500,4500 ,6000,2800,4000}

$\omega_L = 3, \bar{y}_L = 1166.667, SD_L = 531.2459$

$\omega_R = 5, \bar{y}_R = 4160, SD_R = 1078.147$

$$SDR = 1712.409 - \left[\left(\frac{3}{8} * 531.246 \right) + \left(\frac{5}{8} * 1078.147 \right) \right] = 839.35$$

X2<=22.5

L(y):{500,1200,1800,2800}

R(y):{3500,4500, 6000 ,4000}

$\omega_L = 4, \bar{y}_L = 1575, SD_L = 843.727$

$\omega_R = 4, \bar{y}_R = 4500, SD_R = 935.414$

$$SDR = 1712.409 - \left[\left(\frac{4}{8} * 843.727 \right) + \left(\frac{4}{8} * 935.414 \right) \right] = 822.823$$

X2<=27.5

L(y):{500,1200,1800,2800,3500}

R(y):{4500, 6000 ,4000}

$\omega_L = 5, \bar{y}_L = 1960, SD_L = 1069.626$

$\omega_R = 3, \bar{y}_R = 4833.333, SD_R = 849.837$

$$SDR = 1712.409 - \left[\left(\frac{5}{8} * 1069.626 \right) + \left(\frac{3}{8} * 849.837 \right) \right] = 725.204$$

X2<=32.5

L(y):{500,1200,1800,2800,3500,4000}

R(y):{4000, 6000}

$\omega_L = 6, \bar{y}_L = 2383.333, SD_L = 1287.788$

$\omega_R = 2, \bar{y}_R = 5000, SD_R = 1000$

$$SDR = 1712.409 - \left[\left(\frac{6}{8} * 1287.788 \right) + \left(\frac{2}{8} * 1000 \right) \right] = 496.568$$

X2<=37.5

L(y):{500,1200,1800,2800,3500,4500}

R(y):{6000}

$\omega_L = 7, \bar{y}_L = 2614.286, SD_L = 1385.051$

$\omega_R = 1, SD_R = 0$

$$SDR = 1712.409 - \left[\left(\frac{7}{8} * 1385.051 \right) + \left(\frac{1}{8} * 0 \right) \right] = 500.489$$

การแบ่งกลุ่มแรกที่ดีที่สุดคือ $X_1 \leq 10$ และ $X_2 \leq 17.5$

$L(y): \{500, 1200, 1800\}$

$R(y): \{3500, 4500, 6000, 2800, 4000\}$

$\omega_L = 3, \bar{y}_L = 1166.667, SD_L = 531.2459$

$\omega_R = 5, \bar{y}_R = 4160, SD_R = 1078.147$

$$SDR = 1712.409 - \left[\left(\frac{3}{8} * 531.246 \right) + \left(\frac{5}{8} * 1078.147 \right) \right] = 839.35$$

จะได้

กลุ่มข้อมูล(L)

ชั่วโมงที่เล่น(x1)	เลเวล(x2)	คะแนน(y)
2	5	500
5	10	1200
8	15	1800

กลุ่มข้อมูล(R)

ชั่วโมงที่เล่น(x1)	เลเวล(x2)	คะแนน(y)
12	20	2800
15	25	3500
18	35	4000
20	30	4500
25	40	6000

หยุดแบ่ง Node (สร้าง Leaf) ก็ต่อเมื่อ Node นั้นมีข้อมูลน้อยกว่าหรือเท่ากับ 3 ชิ้น

หาจุดแบ่งที่สองที่ดีที่สุดจากกลุ่มข้อมูล(L)

UniqueX1

$$(12+15)/2=13.5$$

$$(15+18)/2=16.5$$

$$(18+20)/2=19$$

$$(20+25)/2=22.5$$

UniqueX2

$$(20+25)/2=22.5$$

$$(25+30)/2=27.5$$

$$(30+35)/2=32.5$$

$$(35+40)/2=37.5$$

$$\bar{y} = \frac{2800 + 3500 + 4000 + 4500 + 6000}{5} = 4160$$

$$SD = \sqrt{\frac{(2800-4160)^2 + (3500-4160)^2 + (4000-4160)^2 + (4500-4160)^2 + (6000-4160)^2}{5}}$$

$$SD = 1078.147$$

$$X1 \leq 13.5$$

$$L(y): \{2800\}$$

$$\omega_L = 1, SD_L = 0$$

$$R(y): \{3500, 4000, 4500, 6000\}$$

$$\omega_R = 4, \bar{y}_R = 4260, SD_R = 1177.455$$

$$SDR = 1078.147 - \left[\left(\frac{1}{5} * 0 \right) + \left(\frac{4}{5} * 1177.455 \right) \right] = 329.816$$

$$X1 \leq 16.5$$

$$L(y): \{2800, 3500\}$$

$$\omega_L = 2, \bar{y}_L = 3150, SD_L = 350$$

$$R(y): \{4000, 4500, 6000\}$$

$$\omega_R = 3, \bar{y}_R = 4833.333, SD_R = 849.837$$

$$SDR = 1078.147 - \left[\left(\frac{2}{5} * 350 \right) + \left(\frac{3}{5} * 849.837 \right) \right] = 428.245$$

X1<=19

L(y):{2800,3500,4000}

R(y):{4500,6000}

$\omega_L = 3, \bar{y}_L = 3433.333, SD_L = 492.161$

$\omega_R = 4, \bar{y}_R = 5250, SD_R = 750$

$$SDR = 1078.147 - \left[\left(\frac{3}{5} * 492.161 \right) + \left(\frac{2}{5} * 750 \right) \right] = 259.578$$

X1<=22.5

L(y):{2800,3500,4000,4500}

R(y):{ 6000}

$\omega_L = 4, \bar{y}_L = 3700, SD_L = 628.49$

$\omega_R = 1, SD_R = 0$

$$SDR = 1078.147 - \left[\left(\frac{3}{5} * 492.161 \right) + \left(\frac{2}{5} * 750 \right) \right] = 575.335$$

X2<=22.5

L(y):{2800}

R(y):{3500,4000,4500,6000}

$\omega_L = 1, SD_L = 0$

$\omega_R = 5, \bar{y}_R = 2760, SD_R = 1177.455$

$$SDR = 1078.148 - \left[\left(\frac{1}{6} * 0 \right) + \left(\frac{5}{6} * 1177.455 \right) \right] = 329.816$$

X2<=27.5

L(y):{2800,3500}

R(y):{4000,4500,6000}

$\omega_L = 2, \bar{y}_L = 3150, SD_L = 350$

$\omega_R = 4, \bar{y}_R = 4833.333, SD_R = 849.837$

$$SDR = 1078.147 - \left[\left(\frac{2}{5} * 350 \right) + \left(\frac{3}{5} * 849.837 \right) \right] = 428.245$$

X2<=32.5

L(y):{2800,3500,4500}

R(y):{4000,6000}

$\omega_L = 3, \bar{y}_L = 3600, SD_L = 697.615$

$\omega_R = 4, \bar{y}_R = 5000, SD_R = 1000$

$$SDR = 1078.147 - \left[\left(\frac{3}{5} * 697.615 \right) + \left(\frac{2}{5} * 1000 \right) \right] = 259.578$$

$$X^2 \leq 37.5$$

$$L(y): \{2800, 3500, 4000, 4500\}$$

$$R(y): \{6000\}$$

$$\omega_L = 4, \bar{y}_L = 3700, SD_L = 628.49$$

$$\omega_R = 1, SD_R = 0$$

$$SDR = 1078.147 - \left[\left(\frac{3}{5} * 492.161 \right) + \left(\frac{2}{5} * 750 \right) \right] = 575.335$$

การแบ่งกลุ่มสองที่ดีที่สุดคือ $X1 \leq 22.5$ หรือ $X2 \leq 37.5$

$L(y): \{2800, 3500, 4000, 4500\}$

$R(y): \{6000\}$

$\omega_L = 4, \bar{y}_L = 3700, SD_L = 628.49$

$\omega_R = 1, SD_R = 0$

$$SDR = 1078.147 - \left[\left(\frac{3}{5} * 492.161 \right) + \left(\frac{2}{5} * 750 \right) \right] = 575.335$$

กลุ่มข้อมูล(L)

ชั่วโมงที่เล่น(x1)	เลเวล(x2)	คะแนน(y)
12	20	2800
15	25	3500
18	35	4000
20	30	4500

กลุ่มข้อมูล(R)

ชั่วโมงที่เล่น(x1)	เลเวล(x2)	คะแนน(y)
25	40	6000

$$\bar{y} = \frac{2800 + 3500 + 4000 + 4500}{4} = 3700$$

$$SD = \sqrt{\frac{(2800 - 3700)^2 + (3500 - 3700)^2 + (4000 - 3700)^2 + (4500 - 3700)^2}{4}}$$

$$SD = 628.49$$

UniqueX1

UniqueX2

$$(12+15)/2=13.5$$

$$(20+25)/2=22.5$$

$$(15+18)/2=16.5$$

$$(25+30)/2=27.5$$

$$(18+20)/2=19$$

$$(30+35)/2=32.5$$

X1<=13.5

L(y):{2800}

$\omega_L = 1, \overline{y}_L = 0$

R(y):{3500,4000,4500}

$\omega_R = 3, \overline{y}_R = 4000, SD_R = 408.3248$

$$SDR = 628.49 - \left[\left(\frac{1}{4} * 0 \right) + \left(\frac{3}{4} * 408.248 \right) \right] = 322.304$$

X1<=16.5

L(y):{2800,3500}

$\omega_L = 2, \overline{y}_L = 3150, SD_L = 350$

R(y):{4000,4500}

$\omega_R = 2, \overline{y}_R = 4250, SD_R = 250$

$$SDR = 628.49 - \left[\left(\frac{2}{4} * 350 \right) + \left(\frac{2}{4} * 250 \right) \right] = 328.49$$

X1<=19

L(y):{2800,3500,4000}

$\omega_L = 3, \overline{y}_L = 3433.333, SD_L = 492.161$

R(y):{4500}

$\omega_R = 1, SD_R = 0$

$$SDR = 628.49 - \left[\left(\frac{3}{4} * 492.161 \right) + \left(\frac{1}{4} * 0 \right) \right] = 259.369$$

X2<=22.5

L(y):{2800}

$\omega_L = 1, SD_L = 0$

R(y):{3500,4000,4500}

$\omega_R = 3, \overline{y}_R = 4000, SD_R = 408.3248$

$$SDR = 628.49 - \left[\left(\frac{1}{4} * 0 \right) + \left(\frac{3}{4} * 408.248 \right) \right] = 322.304$$

X2<=27.5

L(y):{2800,3500}

$\omega_L = 2, \overline{y}_L = 3150, SD_L = 350$

R(y):{4000,4500}

$\omega_R = 2, \overline{y}_R = 4250, SD_R = 250$

$$SDR = 628.49 - \left[\left(\frac{2}{4} * 350 \right) + \left(\frac{2}{4} * 250 \right) \right] = 328.49$$

$$X_2 \leq 32.5$$

$$L(y): \{2800, 3500, 4500\}$$

$$R(y): \{4000\}$$

$$\omega_L = 3, \bar{y}_L = 3433.333, SD_L = 492.161$$

$$\omega_R = 1, \bar{y}_R = 0$$

$$SDR = 628.49 - \left[\left(\frac{3}{4} * 697.615 \right) + \left(\frac{1}{4} * 0 \right) \right] = 105.279$$

การแบ่งกลุ่มสามที่ดีที่สุดคือ $X_1 \leq 16.5$ หรือ $X_2 \leq 27.5$

$$L(y): \{2800, 3500\}$$

$$R(y): \{4000, 4500\}$$

$$\omega_L = 2, \bar{y}_L = 3150, SD_L = 350$$

$$\omega_R = 2, \bar{y}_R = 4250, SD_R = 250$$

$$SDR = 628.49 - \left[\left(\frac{2}{4} * 350 \right) + \left(\frac{2}{4} * 250 \right) \right] = 328.49$$

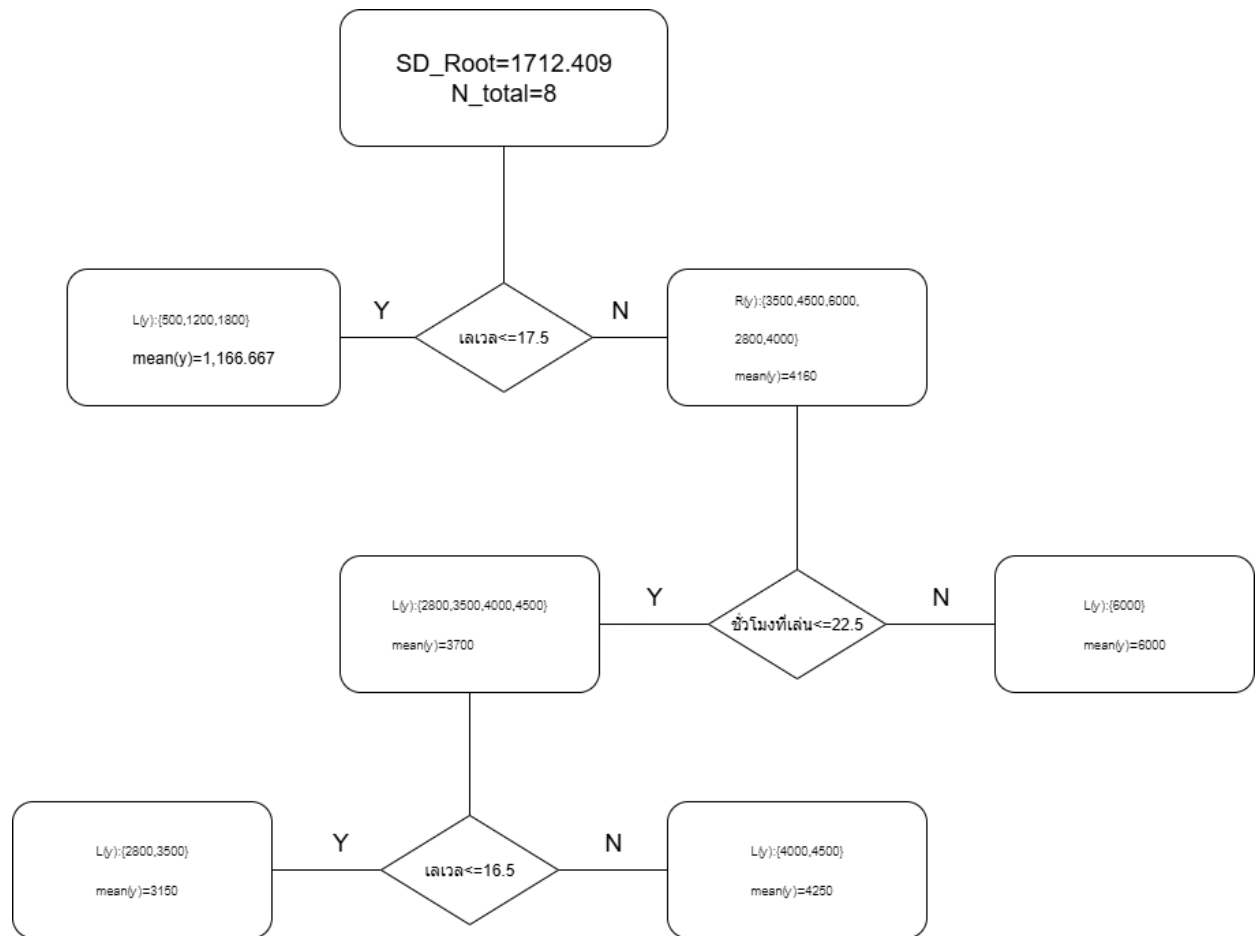
กลุ่มข้อมูล(L)

ชั่วโมงที่เล่น(x1)	เลเวล(x2)	คะแนน(y)
12	20	2800
15	25	3500

กลุ่มข้อมูล(R)

ชั่วโมงที่เล่น(x1)	เลเวล(x2)	คะแนน(y)
18	35	4000
20	30	4500

2. วาดแผนผังต้นไม้ (Decision Tree) ที่สร้างเสร็จแล้ว



3. หากมีผู้เล่นใหม่ที่มีชั่วโมงที่เล่น 10 ชั่วโมง และ เลเวล 18 จงทำนายคะแนนของเขา

∴ หากมีผู้เล่นใหม่ที่มีชั่วโมงที่เล่น 10 ชั่วโมง และ เลเวล 18 จะมีคะแนนอยู่ที่ 4250

โจทย์ข้อที่ 3.1

นักวิเคราะห์หาลินเชื่อมีข้อมูลการอนุมัติสินเชื่อส่วนบุคคล โดยพิจารณาจาก "รายได้ต่อปี (แสนบาท)" (X1) และ "หนี้สินรวม (แสนบาท)" (X2)

ID	รายได้ (X1)	หนี้สิน (X2)	ผลอนุมัติ (Y)
P1	5	1	อนุมัติ
P2	6	3	อนุมัติ
P3	2	2	ไม่อนุมัติ
P4	3	4	ไม่อนุมัติ
P5	7	2	อนุมัติ
P6	4	5	ไม่อนุมัติ

คำสั่ง: ลูกค้าใหม่ (P_new) มีรายได้ 6 แสนบาท และ หนี้สิน 4 แสนบาท จงใช้ K-NN (K=3) ทำนายว่า ลูกค้าคนนี้จะได้รับการอนุมัติหรือไม่?

$$P_1 = \sqrt{(5-6)^2 + (1-4)^2} = 3.162$$

$$P_2 = \sqrt{(6-6)^2 + (3-4)^2} = 1$$

$$P_3 = \sqrt{(2-6)^2 + (2-4)^2} = 4.472$$

$$P_4 = \sqrt{(3-6)^2 + (4-4)^2} = 3$$

$$P_5 = \sqrt{(7-6)^2 + (2-4)^2} = 2.236$$

$$P_6 = \sqrt{(4-6)^2 + (5-4)^2} = 2.236$$

จงใช้ K-NN (K=3) ทำนายว่าลูกค้าคนนี้จะได้รับการอนุมัติหรือไม่?

ID	ระยะห่างจาก P_New	ผลอนุมัติ
P1	1	อนุมัติ
P5,P6	2.236	อนุมัติ/ไม่อนุมัติ
P4	3	ไม่อนุมัติ

∴ P_New อาจจะอนุมัติหรือไม่อนุมัติก็ได้ก็ได้ หรืออิงจากระยะทางเฉลี่ยที่สุด

กลุ่มอนุมัติ P1,P5

$$(1+2.236)/2=1.618$$

กลุ่มไม่อนุมัติ P4,P6

$$(3+2.236)/2=2.618$$

∴ P_New จะจัดอยู่ในกลุ่มอนุมัติ

โจทย์ข้อที่ 3.2

มหาวิทยาลัยแห่งหนึ่งใช้ข้อมูล "เกรดเฉลี่ยตอน ม.ปลาย" (X1) และ "คะแนนสอบเข้า" (X2) เพื่อคัดกรองนักศึกษาที่มีแนวโน้มจะ "เรียนต่อจนจบ" หรือ "ลาออก"

ID	GPA (X1)	คะแนนสอบ (X2)	สถานะ (Y)
S1	3.8	85	เรียนจบ
S2	2.5	60	ลาออก
S3	3.5	90	เรียนจบ
S4	2.8	75	ลาออก
S5	3.2	80	เรียนจบ
S6	2.2	65	ลาออก
S7	3.9	95	เรียนจบ

คำสั่ง: นักเรียนใหม่ (S_{new}) มี GPA 3.0 และ คะแนนสอบ 70 จงใช้ K-NN (K=5) ทำนายสถานะของนักเรียน

คนนี้

$$S_1 = \sqrt{(3.8-3)^2 + (85-70)^2} = 15.021$$

$$S_2 = \sqrt{(2.5-3)^2 + (60-70)^2} = 10.012$$

$$S_3 = \sqrt{(3.5-3)^2 + (90-70)^2} = 20.006$$

$$S_4 = \sqrt{(2.8-3)^2 + (75-70)^2} = 5.004$$

$$S_5 = \sqrt{(3.2-3)^2 + (80-70)^2} = 10.002$$

$$S_6 = \sqrt{(2.2-3)^2 + (65-70)^2} = 5.063$$

$$S_7 = \sqrt{(3.9-3)^2 + (95-70)^2} = 25.016$$

ID	ระยะห่างจากP_New	สถานะ
S4	5.004	ลาออก
S6	5.063	ลาออก
S5	10.002	เรียนจบ
S2	10.012	ลาออก
S1	15.021	เรียนจบ

∴ S_New มีแนวโน้มว่าจะเรียนจบ

โจทย์ข้อที่ 4.1

มีข้อมูล 2 คลาส คือ A (สีฟ้า) และ B (สีแดง)

• คลาส A: P1(2, 5), P2(3, 2)

• คลาส B: P3(6, 4), P4(7, 7)

มีคอนเซนอเส้นแบ่ง (Hyperplane) H1 คือเส้นแนวดิ่ง $x=4.5$

คำสั่ง:

1. จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H1

∴ สมการหลักคือ $1x + 0y - 4.5 = 0$

กลุ่มA

$$P1 = \frac{|2 + 0 - 4.5|}{\sqrt{1^2 + 0^2}} = 2.5$$

$$P2 = \frac{|3 + 0 - 4.5|}{\sqrt{1^2 + 0^2}} = 1.5$$

กลุ่มB

$$P3 = \frac{|6 + 0 - 4.5|}{\sqrt{1^2 + 0^2}} = 1.5$$

$$P4 = \frac{|7 + 0 - 4.5|}{\sqrt{1^2 + 0^2}} = 2.5$$

2. เส้น H1 มี Support Vectors คือจุดใดบ้าง? และมี Margin กว้างเท่าใด?

∴ เส้น H1 มี Support Vectors คือจุด P2 และ P3 และมี Margin = 3

3. จงหาเส้นแบ่งที่ดีที่สุด (Optimal Hyperplane) และ Margin สูงสุดที่เป็นไปได้สำหรับข้อมูลชุดนี้

∴ Xoptimal: $\frac{3+6}{2} = 4.5$ และมี Margin = 3

โจทย์ข้อที่ 4.2

จากข้อมูลชุดเดิมในข้อ 4.1 มีคนเสนอเส้นแบ่งใหม่ H_2 คือ $x+y-8=0$

คำสั่ง:

1. จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H_2

∴ สมการหลักคือ $1x + 1y - 8 = 0$

กลุ่ม A

$$P1 = \frac{|2+5-8|}{\sqrt{1^2+1^2}} = 0.707$$

$$P2 = \frac{|3+2-8|}{\sqrt{1^2+1^2}} = 2.121$$

กลุ่ม B

$$P3 = \frac{|6+4-8|}{\sqrt{1^2+1^2}} = 1.414$$

$$P4 = \frac{|7+7-8|}{\sqrt{1^2+0^2}} = 6.364$$

2. เส้น H_2 มี Support Vectors คือจุดใดบ้าง และ Margin กว้างเท่าใด?

∴ เส้น H_2 มี Support Vectors คือจุด $P1$ และ $P3$ และมี Margin = 2.121

3. เปรียบเทียบกับผลลัพธ์ในข้อ 4.1 เส้น H_2 เป็นเส้นแบ่งที่ดีที่สุดหรือไม่ เพราะอะไร?

∴ เทียบกับ H_1 แล้ว H_2 ไม่ใช่เส้นแบ่งที่ดีที่สุดเพราะ H_1 มี Margin สูงกว่าและมีระยะห่างระหว่างข้อมูลทั้งสองกลุ่มเท่าๆกัน