Nama : M. Faishal Mumtaz

NIM : G64180093

1. a. - Tidak lengkap (missing Value) Pada kolom Height
- Tidak konsisten Pada Satuan kolom weight (kg, lbs)
- Tidak konsisten pada kolom Covid-19 Result
- Tidak konsisten
~~Tidak~~ ~~Noisy~~ Pada kolom Blood group
- Terdapat noisy Pada age ~~Yang~~ ~~dirata~~

Cara mengatasinya :
- menyamakan Satuan Pada kolom yang tidak konsisten
- Pada missing Value dimasukan nilai rata-rata Pada kolom tersebut
- Pada noisy melakukan Pembenaran data.

b. Equal - width (weight)

$w = (max - min) / N$

=> membagi 3 kategori

$(95 - 45) / 3 = 13,3$

Range 1 (rendah) : $45 + 13,3 = 58,3$ (0 - 58,3)

Range 2 (Sedang) : $58,3 + 13,3 = 71,6$ (58,3 - 71,6)

Range 3 (Tinggi) : $71,6 + 13,3 = 84,9 \approx 85$ (71,6 - 85)

2. Single linkege Clustering

| | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ | $T_8$ | $T_9$ | $T_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $J_1$ | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| $J_2$ | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| $J_3$ | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

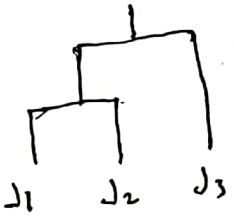
* Sim Jaccard : $\frac{a}{a+b+c}$

Sim $J_1 J_2 : \frac{1}{1+2+5} : \frac{1}{8}$ → karena terkecil maka dipilih

Sim $J_1 J_3 : \frac{2}{2+1+3} = \frac{2}{6}$

Sim $J_2 J_3 : \frac{3}{5+3+2} = \frac{3}{6}$

Dendogramnya =>




3. a(0,1), b(1,1), c(1,2)

* manhattan = $|X_1 - X_0| + |Y_1 - Y_0|$

| | a | b | c |
|---|---|---|---|
| a | - | 1 | 2 |
| b | | - | 1 |
| c | | | - |

* Dist. Jarak terdekat ke-2

dist(a) = c() = 2

dist(b) = (a || c) = 1

dist(c) = (a) = 2

* titik dgn Jarak ≤ dist

$N_2(a) = \{b, c\} = 2$

$N_2(b) = \{a, c\} = 2$

$N_2(c) = \{a, b\} = 2$

* Irdk(a)

$Ird(a) = \frac{2}{1+2} = \frac{2}{3} = 0,66$

$Ird(b) = \frac{2}{2+2} = \frac{2}{4} = 0,5$

$Ird(c) = \frac{2}{2+1} = \frac{2}{3} = 0,66$

* Lofk(a)

$Lof(a) = (0,5 + 0,66) * (1+2) = 3,48$

$LOF(b) = (0,66 + 0,66) * (2+2) = 5,28$

$LOF(c) = (0,66 + 0,5) * (2+1) = 3,48$

∴ Top 1 outlier adalah b dengan

$Lof(b) = 5,28$

4. * Temperatur Badan
   * Normal (36-37)

Temperatur - Normal

| | |
|---|---|
| Ya | 2 |
| Tidak | 2 |

Gini = $1-(2/4)^2-(2/4)^2$
= 0,5

Temperatur - Tidak normal

| | |
|---|---|
| Ya | 3 |
| Tidak | 1 |

Gini = $1-(3/4)^2-(1/4)^2$
= 0,375

- Gini Split
= 4/8 * 0,5 + 4/8 * 0,375
= 0,4375

* Saturasi oksigen
rendah jika ≤ 90%

Saturasi - rendah

| | |
|---|---|
| Ya | 5 |
| tidak | 0 |

Gini = $1-(5/5)^2-(0)^2$
= 0

Saturasi - tinggi

| | |
|---|---|
| Ya | 0 |
| tidak | 3 |

Gini = $1-(0)^2-(3/3)^2$
= 0

Gini split
= 5/8 * 0 + 3/8 * 0
= 0

* Rapid Test - realitif

| | |
|---|---|
| Ya | 3 |
| Tidak | 1 |

Gini = $1-(3/4)^2-(1/4)^2$
= 0,375

Rapid Test - non relatif

| | |
|---|---|
| Ya | 2 |
| Tidak | 2 |

Gini = $1-(2/4)^2-(2/4)^2$
= 0,5

Gini split
= 4/8 * 0,375 + 4/8 * 0,5
= 0,4375

∴ Best Splitnya adalah Saturasi oksigen karena Gini Split yang paling kecil, yaitu 0

5. a

| | | Predicted + | Predicted − | |
|---|---|---|---|---|
| Actual | + | 250 | 250 | 500 |
| Actual | − | 50 | 1450 | 1500 |
| Total | | 300 | 1700 | 2000 |

b. TP = 250

TN = 1450

FP = 50

FN = 250

c. Accuracy = $\frac{250 + 1450}{2000} = 0,85$

Precision = $\frac{250}{250 + 50} = 0,83$

recall = $\frac{250}{250 + 250} = 0,5$

F-measure = $\frac{2 \times 0,83 \times 0,5}{0,83 + 0,5} = 0,624$

d. Model kurang baik dikarenakan nilai recall dan F-measure kecil walaupun tingkat accuracynya tinggi