Jawaban UTS Data Mining

Nama : Berly Setiawan
NIM : G6418004y

(1) a) masalah Data.

① adanya missing value pada atribut height
↳ dapat diatasi dengan mengganti NA tersebut dengan nilai mean karena atribut height berupa numerik.

② ada instance pada atribut weight yang berbeda satuan
↳ dapat dikonversi dari satuan lbs kedalam kg agar sama.

③ Tidak konsisten pada atribut COVID-19 result
↳ dapat diseragamkan saja menggunakan kategori Positive Negative.

④ Noisy data pada atribut Age
↳ dimana hampir tidak mungkin ada orang berusia 350 tahun. Seharusnya dicrosscheck kembali barang kali salah input atau di replace.

⑤ Ada data Dummy pada atribut Nama (Sys(sr789)
↳ dapat di replace dengan isi yang sesuai.

b) Binning atribut Weight (Equal Width.)

- perlu rubah dahulu instance satuan lbs ke kg.

1 lbs = 0,4536 → 120 lbs = 54,43 kg.

maka datanya menjadi :

Weight	kategori
62 kg	sedang
45 kg	rendah
60 kg	sedang
52 kg	rendah
78 kg	berat
54 kg	rendah
48 kg	rendah
54,43 kg	rendah
85 kg	berat
68 kg.	sedang

equal width

$w = \frac{B-A}{n} = \frac{85-45}{3} = 13,33$

Bin 1 → range (45 , 45 + 13,33) = (45 , 58,33)
Bin 2 → range (45+13,33 , 45 + (13,33)2.) = (58,33 , 71,67)
Bin 3 → range (45+(13.33)2 , 85) = (71,67 , 85)

Bin 1	Bin 2	Bin 3
(rendah)	(sedang)	(berat)

② Sim Jaccard + Clustering Single linkage.

$Sim_{jaccard}(J_1, J_2) = \frac{a}{a+b+c} = \frac{1}{1+2+5} = \frac{1}{8} = 0,125$

$Sim_{jaccard}(J_2, J_3) = \frac{a}{a+b+c} = \frac{3}{3+3+2} = \frac{3}{8} = 0,375$

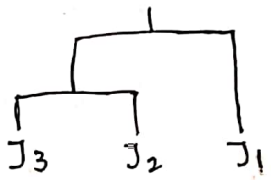
$Sim_{jaccard}(J_1, J_3) = \frac{a}{a+b+c} = \frac{2}{2+1+3} = \frac{2}{6} = 0,333$

Tabel :

	J_1	J_2	J_3
J_1	-	0,125	0,333
J_2		-	0,375
J_3			-

Single linkage mengambil yang terbesar dulu karena $sim_{jaccard}$ adalah ukuran kesamaan.

Dendogram:



③ LOF mencari top 1 outlier (k = 2)

Titik : A (0,1) B (1,1) C (1,2)

Tahap 1 (jarak Manhattan)

$d(a,b) = |0-1| + |1-1| = 1$

$d(a,c) = |0-1| + |1-2| = 2$

$d(b,c) = |1-1| + |1-2| = 1$

	A	B	C
A	-	1	2
B		-	1
C			-

Tahap 2 (jarak tetangga terdekat ke-2 dari titik O)

$dist_2(a) = dist(a,c) = 2$ (c)

$dist_2(b) = dist(b,a) = 1$ (a / c)

$dist_2(c) = dist(c,a) = 2$ (a)

Tahap 3 (hitung $N_k(o)$)

$N_2(A) = \{B, C\}$ karena $dist(A,C), dist(A,B) \leq dist_2(A)$

$N_2(B) = \{A, C\}$ karena $dist(B,A), dist(B,C) \leq dist_2(B)$

$N_2(C) = \{A, B\}$ karena $dist(C,B), dist(C,A) \leq dist_2(C)$.

Tahap 4 (hitung $lrd_k(O)$)

$$lrd_2(A) = \frac{||N_2(A)||}{\text{reach dist}_2(B \leftarrow A) + \text{reach dist}(C \leftarrow A)} = \frac{2}{\max\{1,1\} + \max\{2,2\}} = \frac{2}{1+2} = 0,667$$

$$lrd_2(B) = \frac{||N_2(B)||}{\text{reach dist}_2(A \leftarrow B) + \text{reach dist}(C \leftarrow B)} = \frac{2}{\max\{2,1\} + \max\{2,2\}} = \frac{2}{2+2} = 0,5$$

$$lrd_2(C) = \frac{||N_2(C)||}{\text{reach dist}_2(A \leftarrow C) + \text{reach dist}_2(B \leftarrow C)} = \frac{2}{\max\{2,2\} + \max\{1,1\}} = \frac{2}{2+1} = 0,667$$

Tahap 5 (hitung $LOF_k(O)$)

$$LOF_2(A) = (lrd_2(B) + lrd_2(C)) \times (\text{reach dist}_2(B \leftarrow A) + \text{reach dist}_2(C \leftarrow A))$$
$$= (0,5 + 0,667) \times (1+2) = 3,501$$

$$LOF_2(B) = (lrd_2(A) + lrd_2(C)) \times (\text{reach dist}_2(A \leftarrow B) + \text{reach dist}_2(C \leftarrow B))$$
$$= (0,667 + 0,667) \times (2+2) = 5,336$$

$$LOF_2(C) = (lrd_2(B) + lrd_2(A)) \times (\text{reach dist}_2(B \leftarrow C) + \text{reach dist}_2(A \leftarrow C))$$
$$= (0,5 + 0,667) \times (1+2) = 3,501.$$

Tahap 6 (urutkan LOF)

1) $LOF_2(B) = 5,336$
2) $LOF_2(C) = 3,501$
3) $LOF_2(A) = 3,501$

maka Top 1 outlier adalah titik B (LOF tertinggi).

④ Decision Tree. (gini Index)

① Temperatur → 2 way split berdasarkan informasi yakni normal & tidak normal
(36-37°C) (>37°C)

Temp = normal

Potensi: Ya	2
Potensi: tdk	2
Gini	0,5

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

Temp = tidak normal

Potensi: Ya	3
Potensi: Tdk	1
Gini	0,375

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

gini split Temperature
= $\frac{4}{8} \cdot 0,5 + \frac{4}{8} \cdot 0,375$
= 0,4375.

② Oksigen → 2 way split (rendah = <90%, tidak rendah ≥ 90%)

Oksigen = rendah

Potensi: Ya	5
Potensi: Tdk	0
gini	0

gini = $1-(\frac{5}{5})^2-0^2 = 0$

oksigen = tidak rendah

Potensi: Ya	0
Potensi: Tdk	3
gini	

gini = $1-0^2-(\frac{3}{3})^2 = 0$

gini split oksigen
= $\frac{5}{8} \times 0 + \frac{3}{8} \times 0$
= 0

③ Rapid Test → 2 way split (reaktif & non reaktif)

RT = reaktif

Potensi: Ya	3
Potensi: Tdk	2
Gini	0,48

gini = $1-(3/5)^2-(2/5)^2$
= 0,48

RT = non reaktif

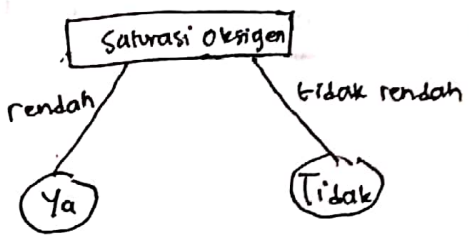
Potensi: Ya	2
Potensi: Tdk	1
Gini	0,444

gini = $1-(2/3)^2-(1/3)^2$
= 0,444

gini split rapid test
= $\frac{5}{8} \times 0,48 + \frac{3}{8} \times 0,44$
= 0,4665

Atribut terbaik sebagai pemisah = Atribut Saturasi Oksigen
· karena nilai gini indexnya paling kecil.

Tree →



(5) Evaluation.

Data aktual
- 500 pasien (+)
- 1500 pasien (-)

a)

aktual \ predicted	Covid = (+)	Covid = (-)
Covid = (+)	250 (TP)	250 (FN)
Covid = (-)	50 (FP)	1450 (TN)

b)
- True Positif (aktual = (+), predicted = (+)) : 250
- True Negatif (aktual = (-), predicted = (-)) : 1450
- False Positif (aktual = (-), predicted = (+)) : 50
- False Negatif (aktual = (+), predicted = (-)) : 250

c) Akurasi = $\frac{TP + TN}{All} = \frac{250 + 1450}{2000} = 0,85$

Precision = $\frac{TP}{TP + FP} = \frac{250}{250 + 50} = 0,833$

Recall = $\frac{TP}{TP + FN} = \frac{250}{250 + 250} = 0,5$

Fmeasure = $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0,833 \times 0,5}{0,833 + 0,5} = 0,6249$

d) Pemodelan yang dilakukan masih belum cukup baik untuk memprediksi kelas positif, dapat dilihat dari nilai Recall yang agak kecil serta nilai F measurenya.

Walau Akurasi yang didapat tinggi tapi model hanya cukup baik untuk memprediksi kelas Negatif. Hal ini juga disebabkan karena data aktual tidak seimbang / imbalance (500 (+) : 1500 (-)).