ASSIGNMENT → IRDM — SHUBHAM SWAMI- A02315672

Q28.14) frequent item set. : | Minimum support value = 0.2 *

Item.	frequency	Support
Milk	5	0.5
Bread.	4.	0.4
Eggs.	4.	0.4.
Juice	3	0.3
butter.	2	0.2
Cookies.	2	0.2
Coffee.	3.	0.3.

since Minimum support val = 0.2 ∴ all items are frequent items as all support values are greater than 0.2.. and. all 1-set items are valid.

2-item set. →	frequency →	support.	
(Milk, Bread)	4	0.4	*
(Milk, eggs).	3	0.3	*
(Milk, Juice)	1.	0.1	
(bread, eggs)	3.	0.3.	*
(bread, cookies).	1	0.1.	
(juice, butter)	1	0.1	
(cookies, butter)	1	0.1	
(Cookies, eggs).	1	0.1	
(coffee , eggs)	1	0.1	
(coffee, Juice)	1	0.1.	

all support values less than 0.2 are not frequent items.

∴

Milk, Bread	0.4
Milk, eggs	0.3
Bread, eggs	0.3.

3-Set items.

	frequency	Support	
Milk - Bread - eggs	3	0.3.	*
Milk - eggs → Bread	0.		
~~Eggs~~ Bread - Eggs - Milk			

∴ 3 item set with support > 0.2 = Milk - Bread - Eggs → 0.3

Q: 28.15 = The above question has only one frequent set of size 3.-

ie Milk eggs Bread with support 70.2.,

Milk bread eggs . = $\frac{\text{Support (Milk, bread, eggs)} = 0.3}{\text{Support (Milk, bread)} = 0.4}$

Milk - bread-eggs = $\frac{3}{4}$ = 0.75

confidence of milk eggs bread = $\frac{\text{Support (M} \cup \text{B} \cup \text{E)}}{\text{Support (M} \cup \text{E)}}$ = $\frac{0.3}{0.3}$ = 1.

Milk → eggs → bread = 1.

Q 28:19 Class Attribute = Repeat customer.

n(Repeat Customer = 'yes') = 7

n(Repeat Customer = 'no') = 3

∴ info gain = $-[3/10 \log_2(3/10) + \frac{7}{10} \log_2(7/10)]$

= 0.88

→ Age attribute

① age (20-30) freq = 5. (4 yes 1 No).

info gain = $-[\frac{4}{5} \log(4/5) + 1/5 \log(1/5)]$ = 0.72

age (31-40) = freq (2) (1 yes 1 No)

Info gain = $-[1/2 \log(1/2) + 1/2 \log(1/2)]$. = 1.

age (41-50) = freq (2)

info gain = $-[2/2 \log 2/2]$ = 0.

age(51-60) = freq (1)

info gain = 0.

$E(\text{Age}) = (0.5 \times 0.72) + (0.2 \times 1) + 0 + 0.$

$= \underline{\underline{0.56}}.$

$\text{Gain}(\text{Age}) = 0.88 - 0.56 = 0.32$

---

city ① NY frequency = 7 (5yes, 2No)

info gain = $-\left[\frac{2}{7}\log(2/7) + \frac{5}{7}\log(5/7)\right] = -\left[2/7(1.79) + 5/7(05)\right)$

$= \underline{\underline{0.86}}$

② LA freq = 2. (1y, 1No)

info gain = $-\left[\log(1/2)\right] = \underline{\underline{1}}.$

③ SF = freq = 1. ∴ Info gain = 0

$E(\text{city}) = (0.7 \times 0.86) + (0.2) + 0 = 0.8$ & $\boxed{\text{Gain} = 0.88 - 0.8 = \underline{\underline{0.08}}}$

Gender: ① F freq (7) (2N 5y)

Infogain = $-1\left[2/7 \log(2/7) + 5/7 \log(5/7)\right] = 0.86.$

② M freq = 3.

Igain = $-1\left[2/3\log(2/3) + 1/3\log(1/3)\right].$

$= \underline{\underline{0.92}}.$

$E(\text{Gender}) = (0.7)(0.86) + (0.3)(0.92) = 0.88$

$\text{Gain}(\text{Gender}) = \underline{\underline{0}}.$

Education:

① College freq = 6. (1, 5)

Igain = $-1\left[\frac{1}{6}\log(1/6) + \frac{5}{6}\log(5/6)\right] = 0.65.$

② Grad = freq = 2

Ig = $\log(1) = 0$

(3) High School freq = 2. ∴ Ig = $\log(2) = 0$

E(education) = $(0.6 \times 0.65) + 0 + 0 = 0.39$

gain (Ed) = 0.49

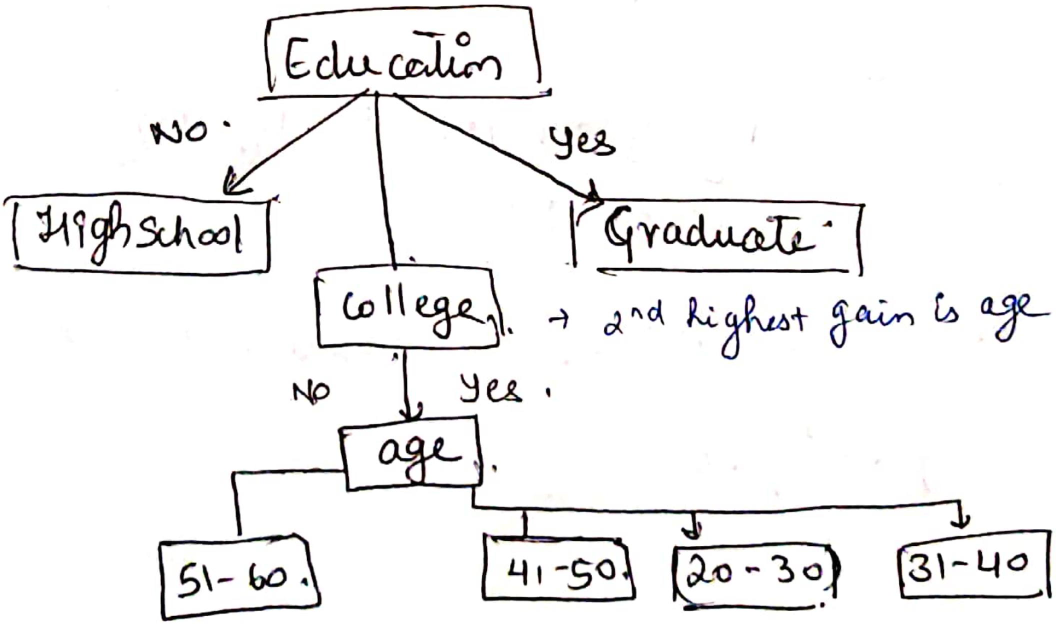
gain(Edu) = 0.49

gain(age) = 0.32

gain(city) = 0.08

gain(gender) = 0

Decision Tree.



Q4 Korth & Silberschatz:

Support (Hammer) = 1/3 = 33% & Support (Nails) = 1/4 = 25%

Rule 1: ∀ Transactions T, true ⇒ buys (T, hammer)

Support = 33% Confidence = 33%

Rule 2: ∀ Transactions T, true ⇒ buys (T, Nails)

Support = 25% Confidence = 25%.

Rule 3 ∀ transactions T, buys(T, ~~[illegible]~~ Hammer) ⇒ buys (T, Nails)

Support = 16.5%, Confidence = $\frac{\text{Support(Rule 3)}}{\text{Support Rule 1}} \times 100 = \frac{\cancel{33}}{2 \times \cancel{33}} \times 100$

Support = 16.5%, Confidence = 50%.

Rule 4: ∀ transactions T buys (T, Nails) ⇒ buys (T, Hammer)

Support = 16.5%, Confidence = $\frac{S(\text{Rule } 4)}{S(\text{Rule } 2)} \times 100 = \frac{33}{\cancel{200}\,2} \frac{\cancel{4}^{2}}{\cancel{4}} \times \cancel{100}$ = 66%.

Q5. Total Documents = 20.

Total Returned Docs = 10 + 8 = 18

Precision = $\frac{\text{Relevant Docs}}{\text{Total Docs Returned}} = \frac{8}{18} = 0.44$.

Recall = $\frac{\text{Relevant Docs}}{\text{Total Docs}} = \frac{8}{20} = 0.4$

Q6. frequency of word in following docs

Doc	computer	Doctoral	Algorithms	Watson
1	2	1	0	0
2	8	0	2	1.
3	20	0	5	0
4.	2	2	0	0
5	~~20~~ 20	0	0	2

IDF = $\log \left(\frac{\text{Total Docs}}{\text{no of docs with term}} \right)$ ⇒ IDF(Computer) = $\log_2. \left(\frac{5}{5} \right) = 0$

IDF (Doctoral) = $\log_2 (5/2) = 1.32$ ✓

IDF (Algorithms) = $\log_2 (5/2) = 1.32$ ✓

IDF (Watson) = $\log_2 (5/2) = 1.32$ ✓

TF = f_i | sum x from | 1 to |V| f_{xj}) TF_{ij} → normalized freq.

f_{ij} → no of occurrences |V| → features

Doc	computer	Doctoral	Algorithm	Watson
1.	2/3 = 0.67	1/3 = 0.33	0	0.
2	8/11 = 0.73	0	2/11 = 0.18	1/11 = 0.09
3	20/25 = ~~0.08~~ 0.8	0	5/25 = 0.2	0
4.	2/4 = 0.5	2/4 = 0.5	0	0.
5	20/22 = 0.91	0	0.	2/22 = 0.09