Question 1

CART Algorithm.

Department	Junior	Senior	Total
Sales	30	80	110
Systems	23	8	31
marketing	4	10	14
Secretary	6	4	10
			165

$$Gini(dept = Sales) = 1 - \left(\frac{30}{110}\right)^2 - \left(\frac{80}{110}\right)^2 = 0.397$$

$$Gini(dept = Systems) = 1 - \left(\frac{23}{31}\right)^2 - \left(\frac{8}{31}\right)^2 = 0.383$$

$$Gini(dept = marketing) = 1 - \left(\frac{4}{14}\right)^2 - \left(\frac{10}{14}\right)^2 = 0.408$$

$$Gini(dept = secretary) = 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.48$$

$$Gini(dept) = \frac{110}{165}(0.397) + \frac{31}{165}(0.383) + \frac{14}{165}(0.408) + \frac{10}{165}(0.48)$$

$$= 0.400$$

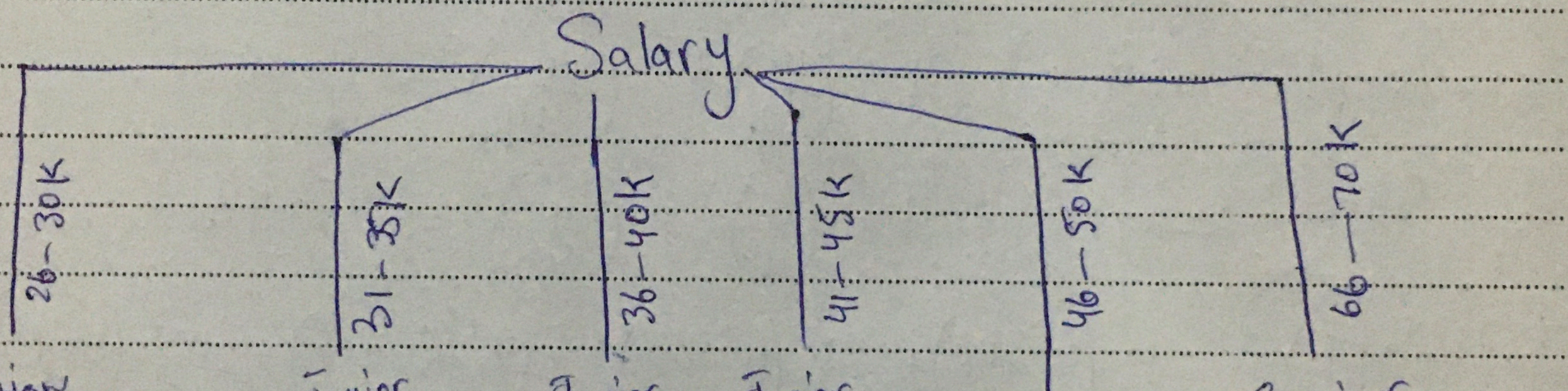
Age	Junior	Senior	Total	Gini
21 — 25	20	0	20	0
26 — 30	49	0	49	0
31 — 35	44	35	79	0.494
36 — 40	0	10	10	0
41 — 45	0	3	3	0
46 — 50	0	4	4	0

$$Gini(Age) = \frac{79}{165}(0.494) = 0.236$$

Salary	Junior	Senior	Total	Gini
26—30K	46	0	46	0
31—35K	40	0	40	0
36—40K	4	0	4	0
41—45K	4	0	4	0
46—50K	23	40	63	0.464
66k—70K	0	8	8	0

Gini (Salary) = $\frac{63}{165}$ (0.464) = 0.177

Splitting on minimum gini value, which is salary



dept	age	status	count
sales	31—35	senior	30
systems	21—25	junior	20
systems	26—30	junior	3
marketing	36—40	senior	10

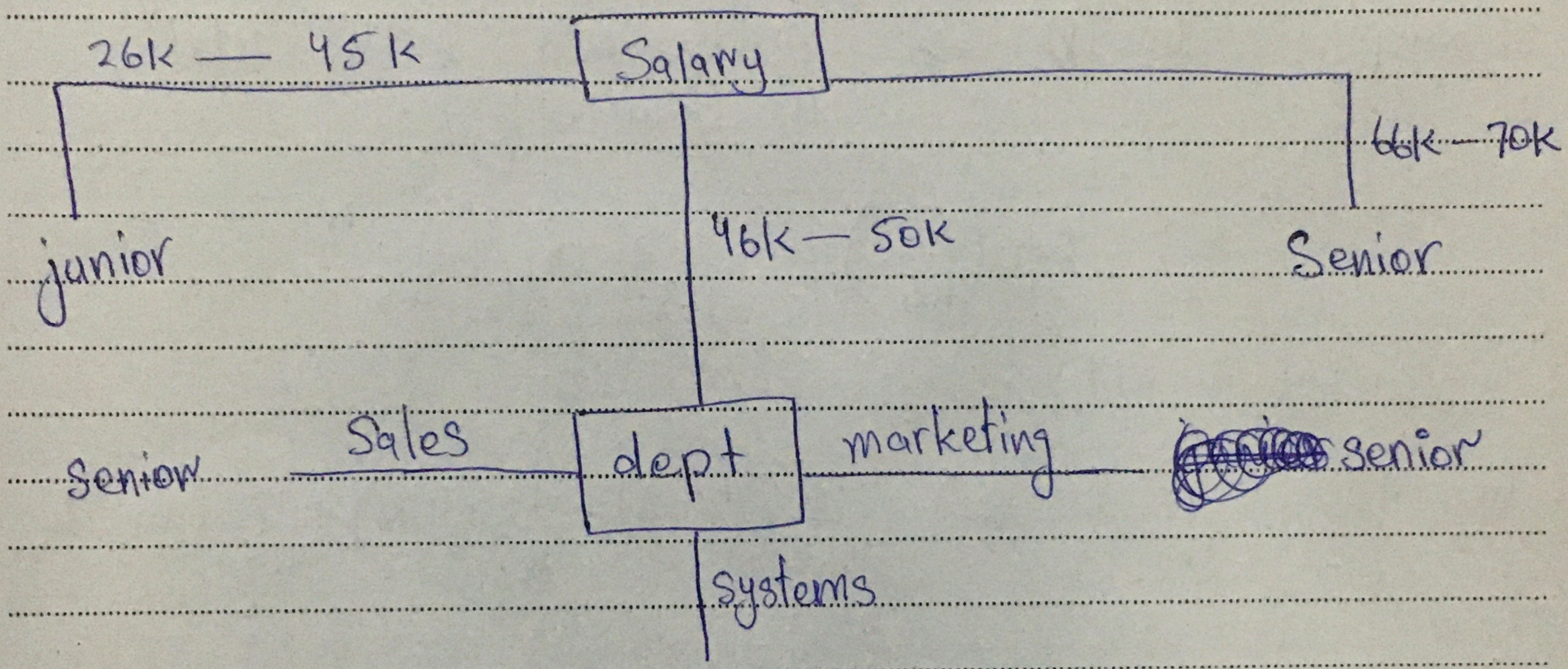
Now, Splitting Second table.

dept	junior	Senior	Total	Gini
Sales	0	30	30	0
Systems	23	0	23	0
marketing	0	10	10	0

Gini (department) = 0

Age	junior	Senior	Total	Gini ~~Count~~
21-25	20	0	20	0
26-30	3	0	3	0
31-35	0	30	30	0
36-40	0	10	10	0

Gini(Age) = 0



★ Both department and age have gini 0, so choosen based on First come basis.

IF — THEN RULE.

IF Salary > 26K AND SALARY < 45K
THEN JUNIOR
IF Salary > 46K AND SALARY < 50K
IF dept = Sales than Senior
if dept = Systems than junior
if dept = marketing than Senior
IF Salary > 66K and Salary < 70K
Than Senior.

C.4.5

Department :

Split info = $-\frac{110}{165}\log_2\left(\frac{110}{165}\right) - \frac{3}{165}\log_2\left(\frac{3}{165}\right) - \frac{14}{165}\log_2\left(\frac{14}{165}\right)$

$-\frac{10}{165}\log_2\left(\frac{10}{165}\right)$

$= 1.388$

Entropy = Info(D) $\Rightarrow$ 52 Seniors, 113 Juniors } 165 total

$= -\frac{52}{165}\log_2\left(\frac{52}{165}\right) - \frac{113}{165}\log_2\left(\frac{113}{165}\right)$

$= 0.894$

Info(A) = $\frac{110}{165}(30,80) + \frac{31}{165}(23,8) + \frac{14}{165}(4,10) + \frac{10}{165}(6,4)$

$= 0.83$

Gain = Info(D) − Info(A)

$= 0.894 - 0.831 = 0.063$.

Gain ratio = $0.063 / 1.388 = 0.045$

Age:

Info(A) = $\frac{20}{165}(20,0) + \frac{49}{165}(49,0) + \frac{79}{165}(44,35) + \frac{10}{165}(0,10)$

$+ \frac{3}{165}(0,3) + \frac{4}{165}(0,4)$

$= 0.47$

Splitting info = $0.3890 + 0.520 + 0.508 + 0.24 + 0.1 + 0.13 = 1.87$

Gain = 0.424

Gain Ratio = $0.424 / 1.87 = 0.224$

Salary

Info(A) = 0.358

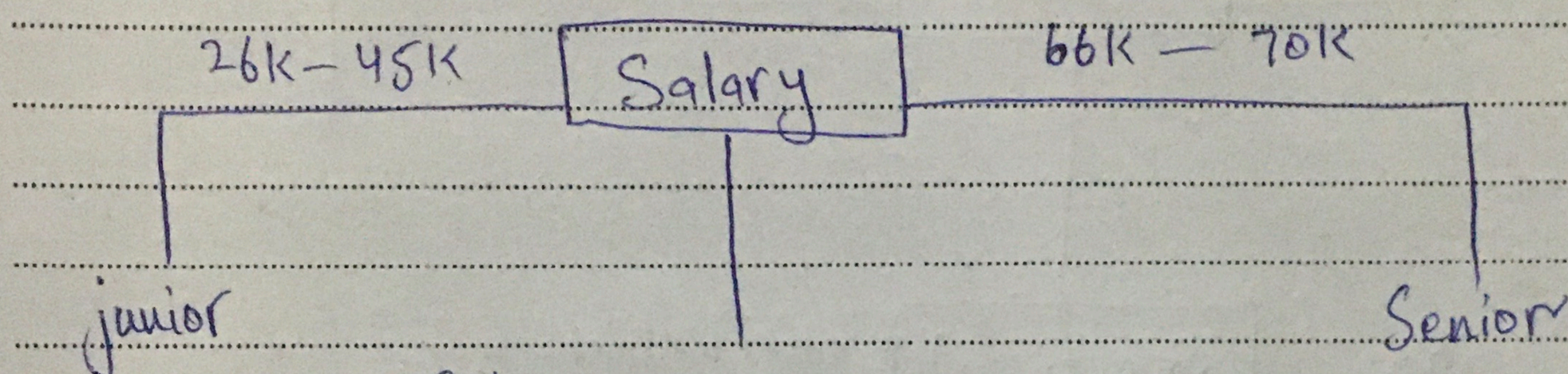
Gain = 0.894 - 0.358

= 0.536

Splitting info = 0.513 + 0.13 + 0.13 + 0.5 + 0.2 + 0.4 = 2.01

Gain Ratio = 0.266

Splitting on highest gain ratio, which is of salary (0.266)



~~Sales~~

dept	Age	Status	Count
sales	31-35	senior	30
systems	21-25	junior	20
systems	26-30	junior	3
marketing	36-40	senior	10

entropy = 2 | 2 | 4

---

department :-

Splitting = $-\frac{30}{63}\log_2\left(\frac{30}{63}\right) - \frac{23}{63}\log_2\left(\frac{23}{63}\right) - \frac{10}{63}\log_2\left(\frac{10}{63}\right)$

= 1.4611

Info(A) = $\frac{30}{63}(0,30) + \frac{23}{63}(23,0) + \frac{10}{63}(0,10)$

= 0

Gain = 1

Gain ratio = $\frac{1}{1.4611}$ = **0.68** ⟶ highest.

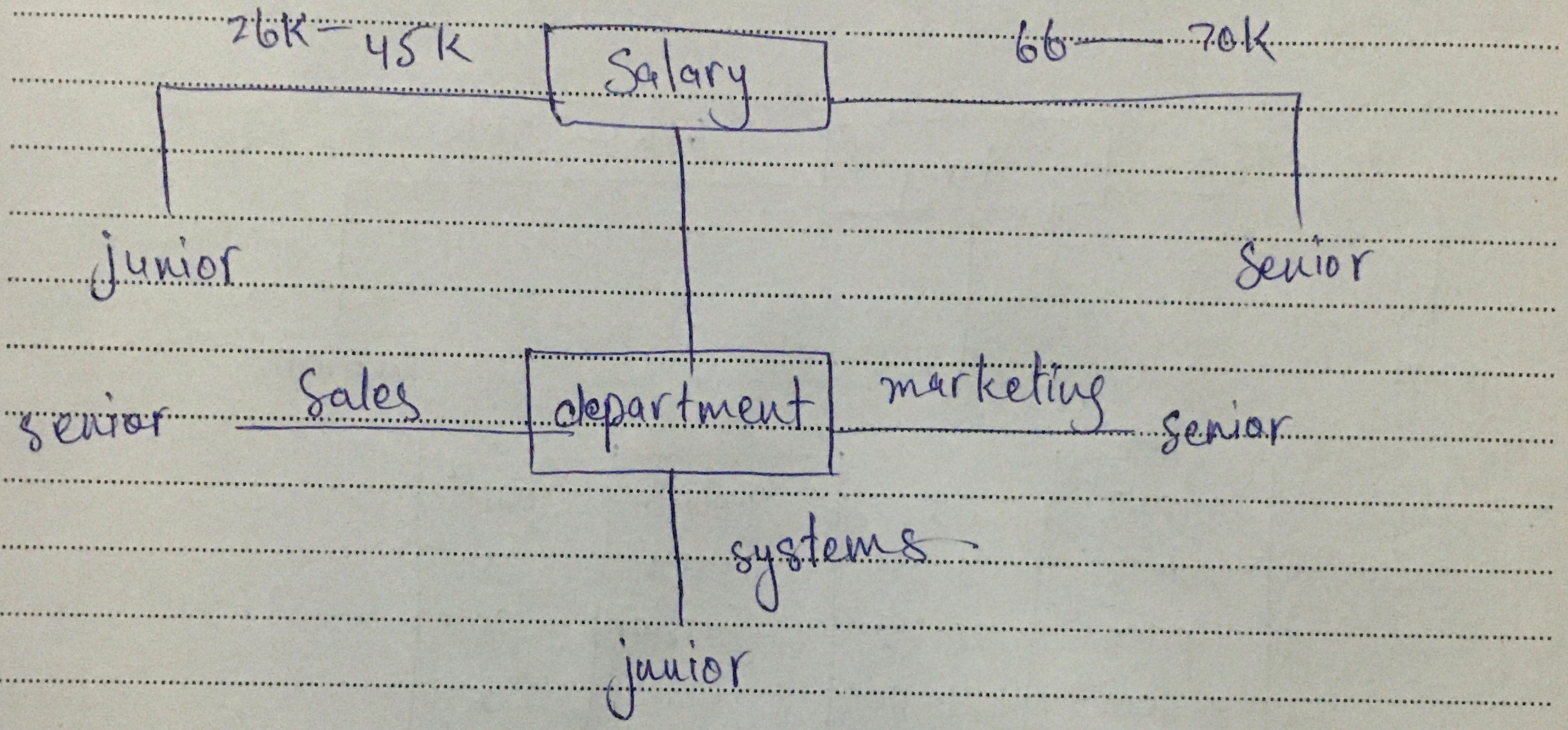
Age :-

Splitting info = $\frac{-20}{63}\log_2\left(\frac{20}{63}\right) - \frac{3}{63}\log_2\left(\frac{3}{63}\right) - \frac{30}{63}\log_2\left(\frac{30}{63}\right)$

= 1.664

info (A) = 6

gain ratio = $\frac{1}{1.644}$ = 0.60



(c) Classification

"systems", 26 — 30, 45 — 60K

Salary ⟶ dept ⟹ junior

So, we get a classification of junior for given ~~data~~ record according to dataset.

Question # 2

Class Label = Purchase,

(A) 150030, Male, 28, 90000, ?

(B) 256899, Female, 40, 60000, ?

Using Male = 0, Female = 1 (ID not included)

Gender, Age, Salary used in Euclidean distance.

$$= \sqrt{(Gender - gender_o)^2 + (Age - Age_o)^2 + (Salary - Salary_o)^2}$$

ID	Gender	Age	Salary	Purchased	Distance A	Distance B
15624510	0	19	19K	0	71K	41K
15810944	0	35	20K	0	70K	40K
15668575	1	26	43K	0	47K	17K
15603246	1	27	57K	0	33K	3K
15804002	0	19	76K	0	14K	16K
15728773	0	27	58K	0	32K	2K
15598044	1	27	84K	0	6K	24K
15694829	1	32	150K	1	60K	90K
15600575	0	25	33K	0	57K	27K
15727311	1	35	65K	0	25K	5K

K = 3

FOR A:

	Distance	Label
1	14000	0
2	6000	0
3	25000	0

major is 0, So label for 150030 is 0

FOR B:

	Distance	Label
1	3000	0
2	2000	0
3	5000	0

major is 0, So label for 256899 is 0.

Class Label : Gender (Male = 0, Female = 1)

Ⓒ 150030, ?, 28, 80000, 1

Ⓓ 256899, ?, 40, 60000, 0

Ⓔ 566989, ?, 15, 20000, 0

Age, Salary, Purchased used in Euclidean Distance.

$$= \sqrt{(Age - Age_o)^2 + (Salary - Salary_o)^2 + (Purchased - purchased)^2}$$

Age	Estimated Salary	Purchased	Gender	C	D	G
19	19K	0	0	71K	41K	1K
35	20K	0	0	70K	40K	20
26	43K	0	1	47K	17K	23K
27	57K	0	1	33K	31K	37K
19	76K	0	0	14K	16K	56K
27	58K	0	0	32K	2K	38K
27	84K	0	1	6K	24K	64K
32	150K	1	1	60K	90K	130K
25	33K	0	0	57K	27K	13K
35	65K	0	1	25K	5K	45K

K = 3.

FOR C :

	Distance	Label
①	14000	0
②	6000	1
③	25000	1

major is 1, So label for 150030 is 1 (Female)

FOR D :

	Distance	Label
①	3000	1
②	2000	1
③	5000	1

major is 1, So label for 256899 is 1 (Female)

FOR E:

	Distance	Label
(1)	1000	0
(2)	20	0
(3)	13000	0

} major is 0. So label for 566989 is 0 (Male)

DISTANCE WEIGHTED KNN

↳ By using $\frac{1}{\text{Euclidean distance}}$.

~~FOR~~ TABLE FOR A, B

Purchased	A	B
0	1.4×10^{-5}	2.44×10^{-5}
0	1.42×10^{-5}	2.5×10^{-5}
0	2.13×10^{-5}	5.88×10^{-5}
0	3.03×10^{-5}	0.00033
0	7.14×10^{-5}	6.25×10^{-5}
0	3.13×10^{-5}	0.0005
0	0.00017	4.17×10^{-5}
1	1.67×10^{-5}	1.11×10^{-5}
0	1.25×10^{-5}	3.73×10^{-5}
0	4×10^{-5}	0.0002

FOR A

$\sum 0 = 0.000407$

$\sum 1 = 0.0000167$

} Since sum for 0 is larger, So Label for 150030 is 0

FOR B

$\sum 0 = 0.00128$

$\sum 1 = 0.000011$

} Since sum for 0 is larger, So Label for 256899 is 0.

TABLE FOR C, D, E

Gender	C	D	E
0	1.4×10^{-5}	2.44×10^{-5}	0.001
0	1.42×10^{-5}	2.5×10^{-5}	0.05
1	2.12×10^{-5}	5.88×10^{-5}	4.35×10^{-5}
1	3.03×10^{-5}	0.00033	2.7×10^{-5}
0	7.14×10^{-5}	6.25×10^{-5}	1.8×10^{-5}
0	3.125×10^{-5}	0.0005	2.6×10^{-5}
1	0.00017	4.17×10^{-5}	1.6×10^{-5}
1	1.67×10^{-5}	1.11×10^{-5}	7.7×10^{-6}
0	1.75×10^{-5}	3.7×10^{-5}	7.7×10^{-5}
1	4×10^{-5}	0.0002	2.2×10^{-5}

FOR C

$\sum 0 = 0.00015$
$\sum 1 = 0.00027$

} Since sum for 1 is large, So label for 150030 is 1 (Female)

FOR D

$\sum 0 = 0.0006489$
$\sum 1 = 0.0006449$

} Since sum for 0 is large, So label for 256899 is 0 (male)

FOR E

$\sum 0 = 0.05$
$\sum 1 = 0.0001$

} Since sum for 0 is large, So label for 566989 is 0 (male)