Homework No. 2 – Basic classification

Course : CSCE 5380.

Raghuram Nimmalapudi
11594739

Q1) a) $E = [5+, 5-]$

$\therefore$ Entrophy $(E) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

$\therefore$ Entrophy $(E) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$

$\therefore$ Entrophy $(E) = 1$

Attribute : Tobacco smoking.

values $\rightarrow$ [Yes, No]

$S_{Yes} \leftarrow [+4, -1]$ $\therefore$ Entrophy $(S_{Yes}) =$

$\Rightarrow -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$

$= 0.7219$

$\therefore S_{No} \leftarrow [+1, -4]$ $\therefore$ Entrophy $(S_{No}) = -\frac{1}{5}\log_2 1/5 - \frac{4}{5}\log_2\frac{4}{5}$

$= 0.7219$

$\therefore$ Gain (S) = Entrophy $(E) = \sum_{v \in V} \frac{|S_v|}{|S|}$ Entrophy (S_v)

$$Gain(S, Tobacco\ smoking) = Entropy(E) - \sum_{v \in (Yes, NO)} \frac{|S_v|}{S} Entropy(S_v)$$

$$= Entropy(E) - \frac{5}{10} Entropy(S_{Yes}) - \frac{5}{10} Entropy(S_{NO})$$

$$= 1 - \frac{5}{10}(0.7219) - \frac{5}{10}(0.7219)$$

$$= 0.2781$$

$\therefore$ Gain (S, Tobacco smoking) = 0.2781

Attribute : Randon Exposure

values $\rightarrow$ [Yes, NO]

$\therefore$ $S_{Yes} \leftarrow [+2, -0]$ $\therefore$ $Entropy(S_{Yes}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}$

$= 0$

$\therefore$ $S_{NO} \leftarrow [+3, -5]$ $\therefore$ $Entropy(S_{NO}) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8}$

$= 0.954$

$\therefore$ $Gain(S, Exposure) = Entropy(E) - \sum_{v \in (Yes, NO)} \frac{|S_v|}{S} Entropy(S_v)$

$= 1 - \frac{2}{10} Entropy(S_{Yes}) - \frac{8}{10} Entropy(S_{NO})$

$= 1 - \frac{2}{10}(0) - \frac{8}{10}(0.954)$

$= 0.2368$

$\therefore$ Gain (S, Exposure) = 0.2368

Attribute: chronic cough

values → [yes, NO]

$S_{yes} \leftarrow [+4, -3]$ ∴ $Entropy(S_{yes}) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}$

$= 0.9852$

$S_{NO} \leftarrow [+1, -2]$ ∴ $Entropy(S_{NO}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9182$

∴ $Gain(S, cough) = Entropy(E) - \sum_{v \in (yes, no)} \frac{|S_v|}{S} Entropy(S_v)$

$= Entropy(E) - \frac{7}{10} Entropy(S_{yes}) - \frac{3}{10} Entropy(S_{NO})$

$= 1 - \frac{7}{10}(0.9852) - \frac{3}{10}(0.9182)$

$= 0.0349$

∴ $Gain(S, cough) = 0.0349$

Attribute: weight loss

values → [yes, NO]

∴ $S_{yes} \leftarrow [+3, -2]$ ∴ $Entropy(S_{yes}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.9709$

∴ $S_{NO} \leftarrow [+2, -3]$ ∴ $Entropy(S_{NO}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.9709$

$\therefore$ Gain (S, weight loss) = Entropy (E) $- \sum_{v \in (yes, no)} \frac{|S_v|}{S}$ Entropy (S_v)

$= $ Entropy (E) $- \frac{5}{10}$ Entropy (S_{yes}) $- \frac{5}{10}$ Entropy (S_{no})

$= 1 - \frac{5}{10}(0.9709) - \frac{5}{10}(0.9709)$

$= 0.0291$

$\therefore$ Gain (S, loss) = 0.0291

Here by gain

Gain (S, Tobacco smoking) = 0.2781

Gain (S, Radon Exposure) = 0.2565

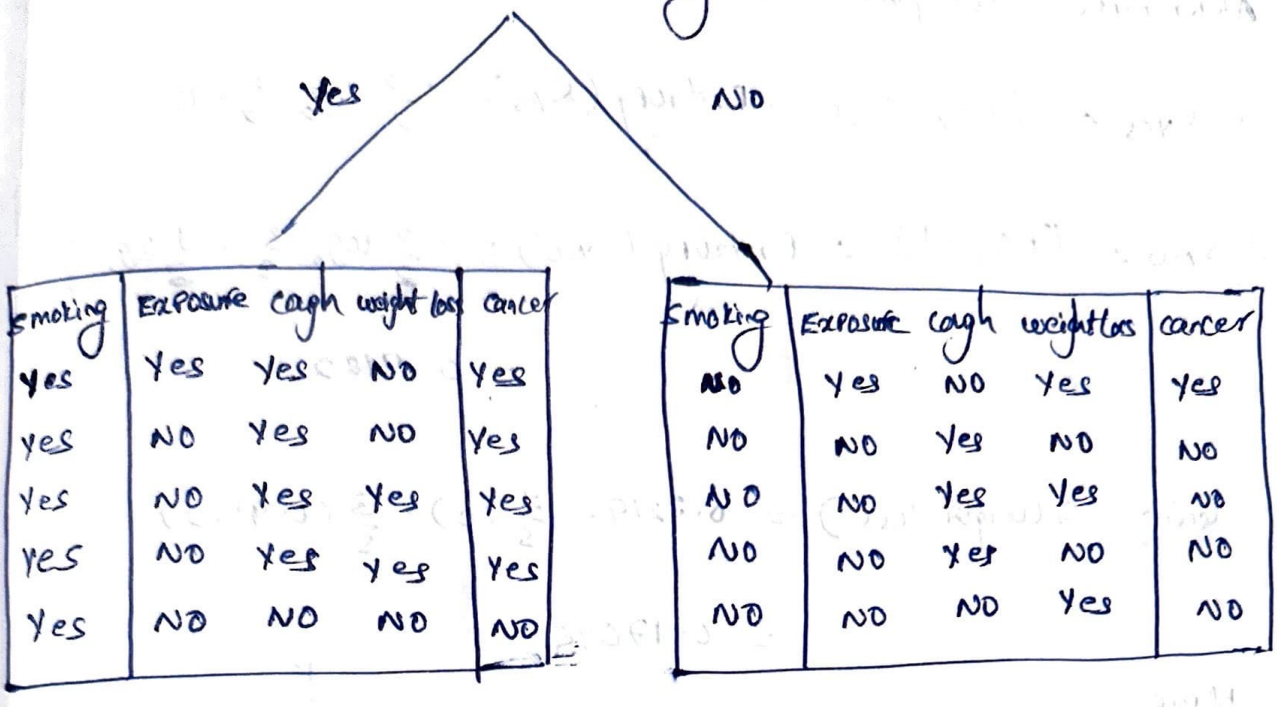
Gain (S, chronic cough) = 0.0349

Gain (S, weight loss) = 0.0291

$\therefore$ In the formation of decision tree, we should consider the maximum information gain.

$\therefore$ So here tobacco smoking as a root.

As there is one "No" for "Yes" in cancer, we have to repeat.



Smoking	Exposure	cough	weight loss	Cancer
Yes	Yes	Yes	No	Yes
Yes	No	Yes	No	Yes
Yes	No	Yes	Yes	Yes
Yes	No	Yes	Yes	Yes
Yes	No	No	No	No

Smoking	Exposure	cough	weightloss	cancer
No	Yes	No	Yes	Yes
No	No	Yes	No	No
No	No	Yes	Yes	No
No	No	Yes	No	No
No	No	No	Yes	No

In "Yes" $\therefore$ Entropy $(E) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.7219$

$\therefore$ ~~Attribute~~ : Exposure

$S_{Yes} \leftarrow [+1, -0]$ $\therefore$ Entropy $(S_{Yes}) = -\frac{1}{1}\log_2\frac{1}{1} = 0$

$= 0$

$S_{NO} \leftarrow [+3, -1]$ $\therefore$ Entropy $(S_{NO}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$

$= 0.8112$

$\therefore$ Gain $(S, \text{Exposure}) = 0.7219 - \frac{1}{5}(0) - \frac{4}{5}(0.8112)$

Gain $(S, \text{Exposure}) = 0.07294$

$\therefore$ Attribute : cough

$\therefore S_{Yes} \leftarrow (+4, -0)$ $\therefore$ Entropy $(S_{Yes}) = -\frac{4}{4}\log_2\frac{4}{4} - 0 = 0$

$\therefore S_{NO} \leftarrow [+0, -1)$ $\therefore$ Entropy $(S_{NO}) = 0 - \frac{1}{1}\log_2\frac{1}{1} = 0$

$\therefore$ Gain $(S, \text{cough}) = 0.7219 - 0 - 0$

$= 0.7219$

Attribute : weight loss

$\therefore S_{yes} \leftarrow [+2, -0]$ $\therefore$ Entropy $(S_{yes}) = -\frac{2}{2}\log_2\frac{2}{2} - 0 = 0$

$\therefore S_{NO} \leftarrow [+2, -1]$ $\therefore$ Entropy $(S_{NO}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

$= 0.9182$

$\therefore$ Gain $(S, \text{weight loss}) = 0.7219 - \frac{2}{5}(0) - \frac{3}{5}(0.9182)$

$= 0.17096$

Here

$\therefore$ Gain $(S, \text{Exposure}) = 0.07294$

$\therefore$ Gain $(S, \text{cough}) = 0.7219$

$\therefore$ Gain $(S, \text{weight loss}) = 0.17096$ [$\therefore$ maximum value of Gain]

$\therefore$ In "NO"

$\therefore$ Entropy $(E) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$

Attribute Exposure:

$\therefore S_{yes} \leftarrow [+1, -0]$ $\therefore$ Entropy $(S_{yes}) = 0$

$\therefore S_{NO} \leftarrow [+0, -4]$ $\therefore$ Entropy $(S_{NO}) = 0$

$\therefore$ Gain $(S, \text{Exposure}) = 0.7219$

Attribute : cough

$\therefore S_{yes} \leftarrow [+0, -3]$ $\therefore$ Entropy $(S_{yes}) = 0$

$\therefore S_{NO} \leftarrow [+1, -1]$ $\therefore$ Entropy $(S_{NO}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= 1$

$Gain(S, cough) = 0.7219 - 0 - \frac{2}{5}(1) = 0.3219$

Attribute : weight loss

$\therefore S_{yes} \leftarrow [+1, -2)$ $\therefore$ Entropy $(S_{yes}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9182$

$\therefore S_{NO} \leftarrow [+0, -2]$ $\therefore$ Entropy $(S_{NO}) = 0$

$\therefore Gain(S, loss) = 0.7219 - \frac{3}{5}(0.9182) - 0$

$\therefore Gain(S, weight\ loss) = 0.1709$

Here

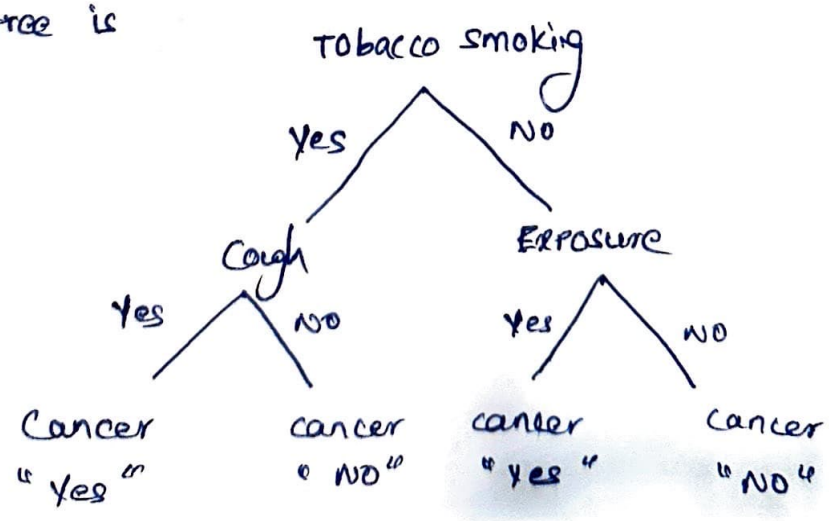
$\therefore Gain(S, Exposure) = 0.7219$

$\therefore Gain(S, cough) = 0.3219$

$\therefore Gain(S, weight\ loss) = 0.1709$

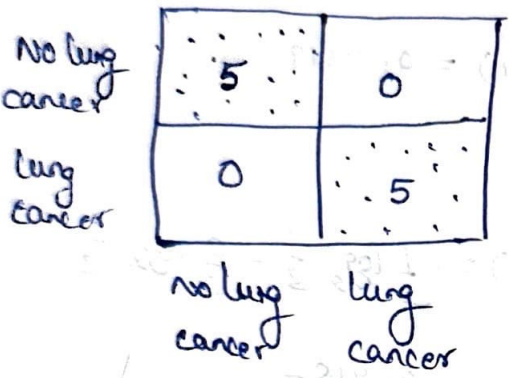
[$\because$ maximum value of Gain]

$\therefore$ The tree is



b) To calculate training error of decision tree we are considering the confusion matrix.

Confusion matrix



Accuracy = TP + TN / TP + TN + FP + FN

= 5 + 5 / 5 + 0 + 5 + 0

= 10 / 10 ⇒ 1

Training Error =

⇒ FP + FN / TP + TN + FP + FN

⇒ 0 + 0 / 0 + 5 + 0 + 5

= 0 / 10

= 0

Q2)

a) The basic decision tree algorithm should be modified as follows to taken into consideration the count of each generalised data tuple.

* The count of each tuple must be integrated into the calculation of the attribute selection measure. (calculate the information gain of each feature).
* Take the count into consideration to determine the most common class among the tuples.
* Considering that all rows don't belong to the same class, split the dataset 's' into subsets using the feature for which the information gain is maximum.
* Make a decision tree node using the feature with the maximum information gain.
* If all rows belong to the same class, make the current node as a leaf node with the class as its label,
* Repeat for the remaining features until we run out of all features, or the decision tree has all leaf nodes.

Q2)

b) the resulting tree is

Salary = 26k..... 30k : junior

= 31k..... 35k : junior

= 36k..... 40k : senior

= 41k..... 45k : junior

= 46k..... 50k (department = secretary : junior

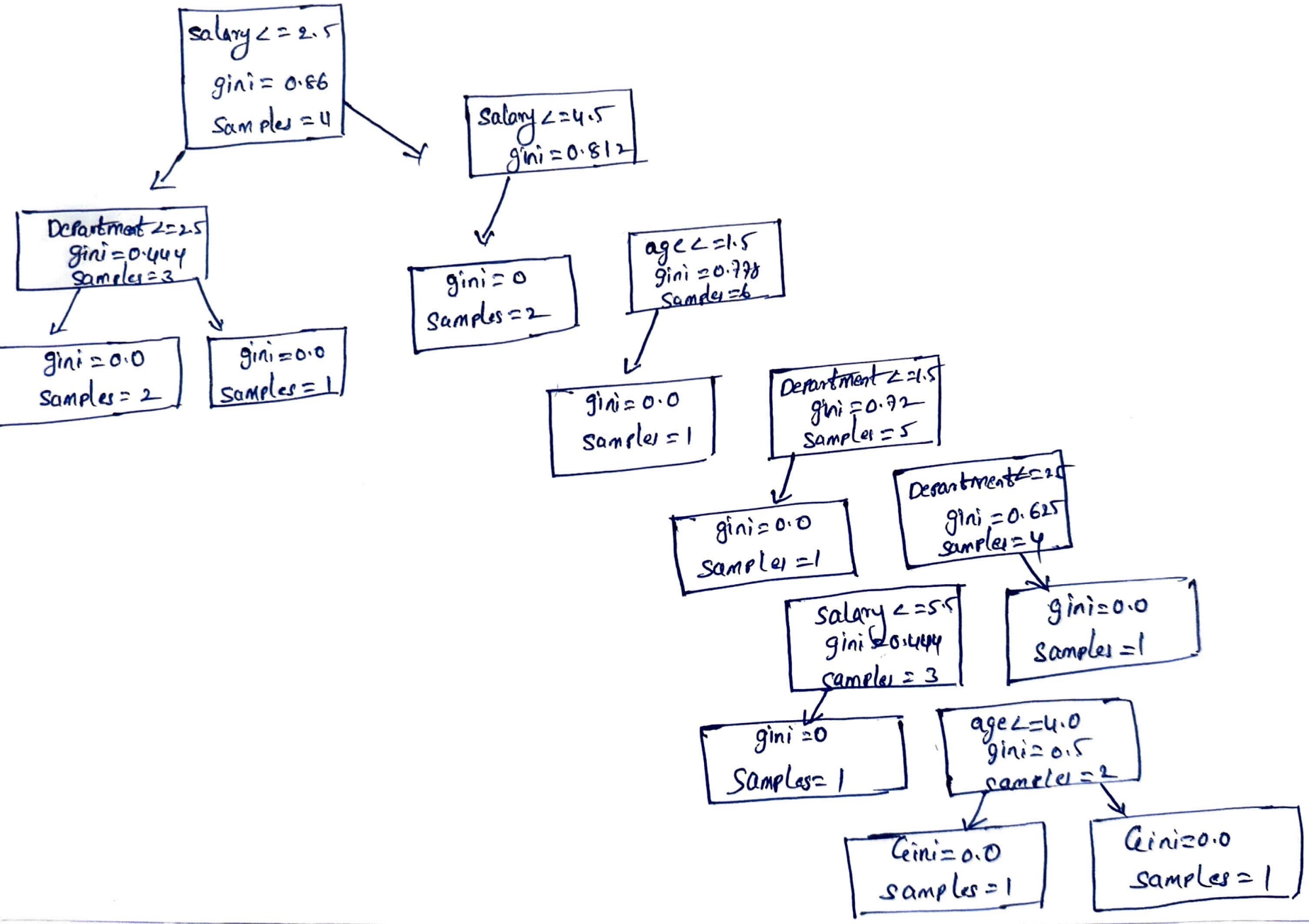
= sales : senior

= systems : junior

= marketing : senior)

= 66k..... 70k: senior.

b) Use code analysis to construct a decision tree from the given data.



c) we first estimate prior probabilities for "status" class labels.

$\therefore P(\text{senior}) = 5/11$

and

$\therefore P(\text{junior}) = 6/11$

Now conditional probabilities

$\therefore P(\text{department}/\text{status})$

class	sales	systems	marketing	secretary
senior	1/5	2/5	1/5	1/5
junior	2/6 = 1/3	2/6 = 1/3	1/6	1/6

P(age/status)

class	21...25	26...30	31...35	36....40	41...45	46....50
senior	0/5 = 0	0/5 = 0	2/5	1/5	1/5	1/5
junior	1/6	3/6 = 1/2	2/6 = 1/3	0/6 = 0	0/6 = 0	0/6 = 0

P(salary/status)

class	26k-30k	21k-35k	36k-40k	41k-45k	46k-50k	66k-70k
senior	0/5 = 0	0/5 = 0	1/5	0/5 = 0	2/5	2/5
junior	2/6 = 1/3	1/6 = 0	0/6 = 0	1/6	2/6 = 1/3	0/6

∴ so for test instance

$$\therefore V_{NB} = \underset{v_j \in [Yes, No]}{\operatorname{argmax}} \; P(v_j) \prod_i P(a_i | v_j)$$

∴ (systems, "26 .... 30", "46k - 50k")

Two labels → Senior, Junior

∴ P(senior|a) = P(senior) * P(systems/senior) * P(26 .... 30/senior) + P(46k-50k/senior)

$$= \frac{5}{11} * \frac{2}{5} * 0 * \frac{2}{5}$$

$\therefore P(\text{senior}/a) = 0$

$\therefore P(\text{Junior}/a) = P(\text{Junior}) * P(\text{systems}/\text{junior}) *$
$P(26 \ldots 30 / \text{Junior}) * P(46k - 50k / \text{Junior})$

$= \frac{6}{11} * \frac{1}{3} * \frac{1}{2} * \frac{1}{3}$

$\therefore P(\text{Junior}/a) = 0.030$

$\therefore$ by comparing the two values and by the maximum values was Junior.

$\therefore$ so
Hence, the label for this instance (person) is "Junior"

$\therefore$ since $P(\text{senior}/a) < P(\text{Junior}/a)$.