1. **Suppose that we have age data including the following numbers in sorted order. Then answer the questions below.**

age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

(a) Use smoothing by <u>Bin Means</u> to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

Step 1: Order Data if not already ordered.

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Step 2: Bin ordered values of depth 3:

[13, 15, 16], [16, 19, 20], [20, 21, 22], [22, 25, 25], [25, 25, 30], [33, 33, 35], [35, 35, 35], [36, 40, 45], [46, 52, 70]

Step 3: Compute the mean of each bin:

15, 18, 21, 24, 27, 34, 35, 40, 56

Step 4: Replace each value in the bin with the computed mean:

[15, 15, 15],[18, 18, 18],[21, 21, 21],[24, 24, 24],[27, 27, 27],[34, 34, 34],[35, 35, 35],[40, 40, 40],[56, 56, 56]

Effect: Smoothing by bin means discretizes an ordered set of data and reduces the noise in a data set by replacing the values in a set with the mean value.

(b) Use <u>IQR measure</u> to determine if there are any outliers in this data.

Step 1: Order Data if not already ordered.

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Step 2: Calculate Q1

27 values/4 = 6.75 = 7

Q1 = 7^{th} Position = 20

Step 3: Calculate Q3

Q3 = 7*3 = 21^{st} Positon = 35

Step 3: Calculate Interquartile Range (IQR)

IQR = Q3 - Q1

= 35 - 20 = 15

Step 4: Calculate Lower Bound

Lower bound = Q1 – (1.5 * IQR)

= 20 – (1.5 * 15) = -2.5

Step 5: Calculate Upper Bound

Upper bound = Q3 + (1.5 * IQR)

= 35 + (1.5 * 15) = 57.5

Step 6: Determine Outliers

Any value below -2.5 or above 57.5 would be considered an outlier. Therefore 70 is an outlier.

(c) Use Min-Max Normalization to transform the value 35 for age onto the range [0.0, 1.0].

Min-max normalization is also referred to as feature scaling.

transform the value 35 onto the range [0.0, 1.0],

Step 1: Order Data if not already ordered.

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Step 2: Use the formula:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Step 3: Identify the X values:

X_{min}=13 X_{max}=70 X = 35

Step 4: Calculate the feature scale.

X_{scaled} = (35 - 13) / (70 - 13) = 22/57 = .386

(d) Use Z-Score Normalization to transform the value 35 for age? (you need to compute mean and standard deviation first)

Step1: Calculate Mean

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{\sum x_i}{n}$$

∑(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)/27

809/27 = 29.96 =30

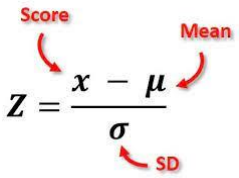
Step2: Calculate Standard Deviation

$$\text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

N-1	26
Sum	809
Mean	30
Var	167
Std Dev	13

Step 3: Calculate Z-Score



(35 – 30)/13 = .38

(e) Use normalization by <u>Decimal Scaling</u> to transform the value 35 for age.

Step 1: Order Data if not already ordered.

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Step 2: Use the Decimal Scaling Formula

$$v' = \frac{v}{10^j}.$$

V = value to to normalize

J = number of digits in the greatest value of ordered list

Step 3: Calculate the Normalize Value

$35/10^2 = 35/100 = .35$

department	age	salary	status	count
sales	31_35	46K_50K	senior	30
sales	26_30	26K_30K	junior	40
sales	31_35	31K_35K	junior	40
systems	21_25	46K_50K	junior	20
systems	31_35	66K_70K	senior	5
systems	26_30	46K_50K	junior	3
systems	41_45	66K_70K	senior	3
marketing	36_40	46K_50K	senior	10
marketing	31_35	41K_45K	junior	4
secretary	46_50	36K_40K	senior	4
secretary	26_30	26K_30K	junior	6

Table 1: Data shows the count of each feature combination. For instance, there are 30 senior sales staff who are 31...35 years old and have 46...50K salary. Since each combination is unique, their corresponding groups are mutually exclusive which implies counts are not double counts for any of the cases. Notice that the status column is the class label to indicate whether someone is a junior or senior.

department	age	salary	status	count
sales	31_35	46K_50K	senior	30
sales	26_30	26K_30K	junior	40
sales	31_35	31K_35K	junior	40
systems	21_25	46K_50K	junior	20
systems	31_35	66K_70K	senior	5
systems	26_30	46K_50K	junior	3
systems	41_45	66K_70K	senior	3
marketing	36_40	46K_50K	senior	10
marketing	31_35	41K_45K	junior	4
secretary	46_50	36K_40K	senior	4
secretary	26_30	26K_30K	junior	6

Gain (Dept)						
Dept	senior (p)	junior (n)	total	Info(D)	$INFO_{Dept}(D)$	Gain(Dept)
marketing	10	4	14	0.863121	0.073234472	
sales	30	80	110	0.845351	0.563567291	
secretary	4	6	10	0.970951	0.058845491	
systems	8	23	31	0.823812	0.154776731	
total	52	113	165	0.899031	0.850423985	0.05

Gain (Age)						
Age	senior (p)	junior (n)	total	Info(D)	$INFO_{Age}(D)$	Gain(Age)
21_25	0	20	20	0	0	
26_30	0	49	49	0	0	
31_35	35	44	79	0.990617	0.47429565	
36_40	10	0	10	0	0	
41_45	3	0	3	0	0	
46_50	4	0	4	0	0	
total	52	113	165	0.899031	0.47429565	0.42

Gain (Salary)						
Salary	senior	junior	total	Info(D)	$INFO_{Salary}(D)$	Gain(Salary)
26K_30K	0	46	46	0	0	
31K_35K	0	40	40	0	0	
36K_40K	4	0	4	0	0	
41K_45K	0	4	4	0	0	
46K_50K	40	23	63	0.946819	0.361512645	
66K_70K	8	0	8	0	0	
total	52	113	165	0.899031	0.361512645	0.54

3. **Using information gain on the data in Table 1, do calculations for two levels of a decision tree which decides whether a person is senior or junior. Please show your calculations and clearly write down your junior and senior counts not to confuse yourself. Note that you need to calculate the information gain for all attributes (department, age, salary) and pick the one to start your tree. In your subsets of your data, you'll perform the same operation for the attributes available. (You can use a computing environment to write the mathematical expressions. i.e., p∗logp; (1/2)∗log2(1/2))**

- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

4. **Using the decision tree you generate if-then rules.**

Step 1: Rank Order Gain Values

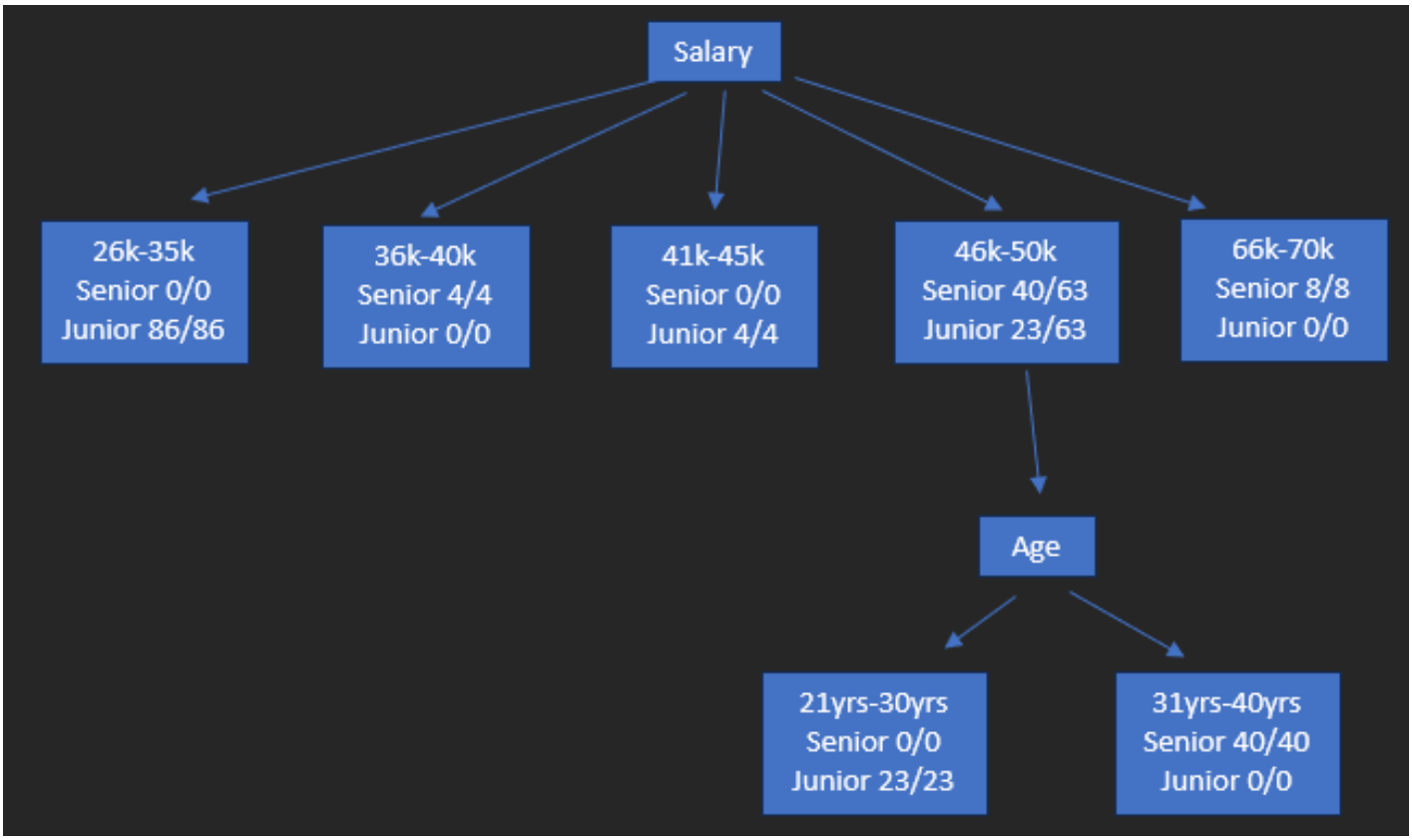
1st Gain (Salary) = 0.54

2nd Gain (Age) = 0.42

3rd Gain (Dept) = 0.05

Step 2: Calculate age subset of non-leaf node(s) and build Tree

Age	senior	junior	total
21_25	0	20	20
26_30	0	3	3
31_35	30	0	30
36_40	10	0	10
Grand Total	40	23	63



Step 3: Calculate if then rules

If Salary between (26k-36k) or (41k-45k) then “Junior”

If Salary between (36k-40k) then “Senior”

If Salary between (41k-45k) then “Junior”

If Salary between (46k,-50k) and Age between (21yrs-30yrs) then “Junior”

If Salary between (46k,-50k) and Age between (31yrs-40yrs) then “Senior”

If Salary between (66k-70k) then “Senior”