HW3 Mike Turley CSC587-W1

1. Data Preprocessing on the Age List

Ages (sorted):
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) Smoothing by Bin Means (bin depth = 3)
There are 27 values ⇒ 9 bins of size 3. Compute each bin's mean and replace members by that mean.

B1: [13,15,16] → 14.67
B2: [16,19,20] → 18.33
B3: [20,21,22] → 21.00
B4: [22,25,25] → 24.00
B5: [25,25,30] → 26.67
B6: [33,33,35] → 33.67
B7: [35,35,35] → 35.00
B8: [36,40,45] → 40.33
B9: [46,52,70] → 56.00

Smoothed sequence:
14.67, 14.67, 14.67, 18.33, 18.33, 18.33, 21, 21, 21, 24, 24, 24, 26.67, 26.67, 26.67, 33.67, 33.67, 33.67, 35, 35, 35, 40.33, 40.33, 40.33, 56, 56, 56

Comment: Smoothing reduces local noise and flattens within-bin variation. It pulls extremes toward their bin's center, preserving trends but losing fine detail.

(b) Outliers via IQR
Q1 = 20.5, Q3 = 35.0 ⇒ IQR = 14.5

Lower fence = Q1 − 1.5×IQR = −1.25
Upper fence = Q3 + 1.5×IQR = 56.75
Values > 56.75 are outliers ⇒ 70 is an outlier.

(c) Min–max normalize age=35 to [0,1]
min=13, max=70:
(35 − 13)/(70 − 13) = 0.386

(d) z-score normalize age=35
Mean ≈ 29.96; population std ≈ 12.70
z = (35 − 29.96)/12.70 ≈ 0.40

(e) Decimal scaling
Max = 70 ⇒ divide by 100 ⇒ 35 → 0.35

2. Function for General Min–Max Normalization
Formula:

$x' = new\_min + (x - old\_min)/(old\_max - old\_min) \times (new\_max - new\_min)$

This rescales any variable to a desired range such as [0,1] or [5,10].

3. Two-Level Decision Tree using Information Gain

Dataset includes attributes: department, age, salary, and class (status).

Root Entropy:
P(junior)=113/165, P(senior)=52/165 ⇒ H(Y)=0.899

Compute IG for each attribute:

IG(department) ≈ 0.049
IG(age) ≈ 0.425
IG(salary) ≈ 0.538 ← Highest

Root split = salary.

For salary=46–50K (mixed branch: 23 junior, 40 senior; H≈0.947):
Splitting by department (or age) gives pure leaves.

Final tree:

Root: salary
- 26–30K → junior
- 31–35K → junior
- 36–40K → senior
- 41–45K → junior
- 46–50K → split by department
  - sales → senior
  - systems → junior
  - marketing → senior
- 66–70K → senior

4. If–Then Rules

1. IF salary ∈ 26–30K THEN status = junior
2. IF salary ∈ 31–35K THEN status = junior
3. IF salary ∈ 36–40K THEN status = senior
4. IF salary ∈ 41–45K THEN status = junior
5. IF salary ∈ 66–70K THEN status = senior
6. IF salary ∈ 46–50K AND department = sales THEN status = senior
7. IF salary ∈ 46–50K AND department = systems THEN status = junior
8. IF salary ∈ 46–50K AND department = marketing THEN status = senior

This two-level decision tree provides 100% classification purity.