Name: Answer Key

After the first exam in a data mining course, the results of the exam were recorded along with some information about each student. The data is below:

ID	Passed All Assignments	GPA	Language	Passed Exam
1	No	3.1	Python	Yes
2	No	2.0	Python	No
3	Yes	3.5	C++	Yes
4	Yes	2.5	Java	No
5	Yes	3.9	Python	No
6	No	3.3	C++	Yes
7	Yes	3.2	Java	Yes

We want to use the above data to create a decision tree that can predict which students will pass the exam.

What is the class label? passed exam

What are the attributes? passed all assignments, GPA, language

In order to create a decision tree, we need to decide which attribute to split on first. To do this, we must calculate the **gain** of splitting on each of our attributes.

1. Start by calculating the impurity of the parent. Use entropy as the measure of impurity.

$$\text{Entropy(parent)} = -\left(\tfrac{4}{7}\right)\log_2\left(\tfrac{4}{7}\right) - \left(\tfrac{3}{7}\right)\log_2\left(\tfrac{3}{7}\right)$$

$$= 0.985$$

Name: _____

2. Next, calculate the information gain of splitting on 'Passed All Assignments'. The gain of splitting on 'Passed All Assignments' will be the entropy of the parent minus the entropy of making this split. (We want to know how much the impurity decreases by making this split.)

y: 2, N: 2 y: 2, N: 1

Entropy(Y) = 1 Entropy(N) = $-\frac{2}{3}\log(\frac{2}{3}) - \frac{1}{3}\log(\frac{1}{3}) = 0.918$

Entropy(split) = $(\frac{4}{7})(1) + (\frac{3}{7})(0.918) = 0.965$

$Gain_{split} = 0.985 - 0.965 = 0.02$

3. Next, calculate the information gain of splitting on 'Language'. The gain of splitting on 'Language' will be the entropy of the parent minus the entropy of making this split. (We want to know how much the impurity decreases by making this split.)

y: 1 N: 2

Entropy(Python) = $-\frac{1}{3}\log(1/3) - \frac{2}{3}\log(2/3) = 0.918$

y: 2 N: 0

Entropy(C++) = 0

y: 1 N: 1

Entropy(Java) = 1

$Entropy_{split} = (\frac{3}{7})(0.918) + (\frac{2}{7})(0) + (\frac{2}{7})(1) = 0.679$

$Gain_{split} = 0.985 - 0.679 = 0.306$

Name: _____

4. Next, calculate the information gain of splitting on 'GPA'. Because GPA is a continuous attribute, we need to try different candidate threshold values. Determine all of the candidate thresholds, then calculate the information gain for each of them.

No	No	Yes	Yes	yes	yes	No
2.0	2.5	3.1	3.2	3.3	3.5	3.9

① splitting at 2.8

≤2.8: 2

>2.8: 5

Entropy (≤2.8) = 0

Entropy (>2.8) = $-\frac{4}{5}\log(\frac{4}{5}) - \frac{1}{5}\log(\frac{1}{5}) = 0.722$

Entropy split = $\frac{2}{7}(0) + \frac{5}{7}(0.722) = 0.516$

Gain split = 0.985 - 0.516 = 0.469

② splitting at 3.7

≤ 3.7 : 6

> 3.7 : 1

Entropy (≤3.7) = $-\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6} = 0.918$

Entropy (>3.7) = 0

Entropy split = $\frac{6}{7}(0.918) + \frac{1}{7}(0) = 0.787$

Gain split = 0.985 - 0.787 = 0.198

↓

better split threshold

5. Impurity metrics (like Gini and entropy) favor attributes with more values. Because 'Language' can be split 3 ways, but 'Passed All Assignments' and 'GPA' are only split 2 ways, we should use **gain ratio** to compare the attributes, rather than just gain.

   Calculate the **split info** for each of the three attributes. With that, calculate the **gain ratio** for each of the three attributes.

Splitinfo (passed all assignments) = $-(\frac{4}{7})\log(\frac{4}{7}) - (\frac{3}{7})\log(\frac{3}{7}) = 0.985$

splitinfo (Language) = $(\frac{3}{7})\log(\frac{3}{7}) - (\frac{2}{7})\log(\frac{2}{7}) - (\frac{2}{7})\log(\frac{2}{7}) = 1.557$

splitinfo (GPA, 2.8) = $-\frac{2}{7}\log(\frac{2}{7}) - \frac{5}{7}\log(\frac{5}{7}) = 0.863$

splitinfo (GPA, 3.7) = $-\frac{6}{7}\log(\frac{6}{7}) - \frac{1}{7}\log(\frac{1}{7}) = 0.58$

Gain ratio (passed all assignments) = $\frac{0.02}{0.985} = 0.02$

Gain ratio (Language) = $\frac{0.306}{1.557} = 0.197$

Gain ratio (GPA, 2.8) = $\frac{0.469}{0.863} = 0.543$

Gain ratio (GPA, 3.7) = $\frac{0.198}{0.58} = 0.33$

6. Based on gain ratio, which attribute do you choose to be the first node in the tree?

Based on gain ratio, I would choose GPA with a threshold of 2.8 as it has the highest Gain ratio of the three.