Name: Dhruv Arora

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 and choose the one with the highest gain.

1. Start by calculating the impurity of the parent. (At this point, the parent is the whole dataset.) Use Gini as the measure of impurity.

$$\text{Gini} = 1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right)$$

$$= 1 - (0.327 + 0.184)$$

$$= \boxed{0.490}$$

Yes	No
4	3

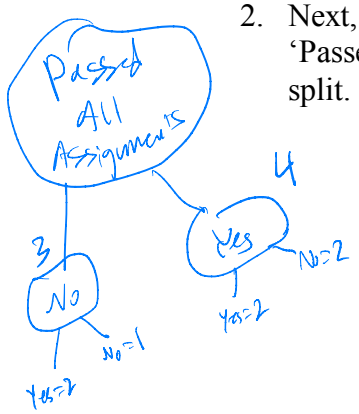
total = 7

Name: ____

(Blank page for printing)

Name: _____

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



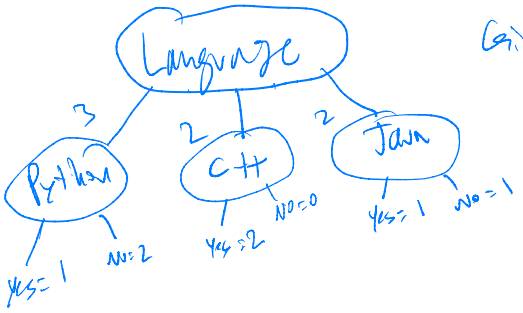
$$Gini(split) = \frac{3}{7}\left(1-\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right)\right) + \frac{4}{7}\left(1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)\right)$$

$$= 0.190 + 0.286$$

$$Gini(split) = \boxed{0.476}$$

$$Gain = 0.490 - 0.476 = \boxed{0.014}$$

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



$$Gini(split) = \frac{3}{7}\left(1-\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right)\right) + \frac{2}{7}\left(1-\left(\left(\frac{2}{2}\right)^2-(0)^2\right)\right) + \frac{2}{7}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]$$

$$= 0.190 + 0 + 0.143$$

$$= \boxed{0.334}$$

$$Gain = 0.490 - 0.334 = \boxed{0.156}$$

4. Next, calculate the Gini of splitting on 'GPA'. Because GPA is a continuous attribute, we need to try different candidate split-point values. Determine the candidate split-points, then calculate the Gini for each of them, and the gain for each of them.

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

2.8 → $Gini = \frac{2}{7}(0) + \frac{5}{7}\left(1-\left(\left(\frac{4}{5}\right)^2+\left(\frac{1}{5}\right)^2\right)\right)$

$= \boxed{0.229}$ } Better

$Gain = 0.490 - 0.229 = \boxed{0.261}$

3.7 → $Gini = \frac{1}{7}(0) + \frac{6}{7}\left(1-\left(\left(\frac{4}{6}\right)^2+\left(\frac{2}{6}\right)^2\right)\right)$

$= \boxed{0.381}$

$Gain = 0.490 - 0.381 = \boxed{0.109}$

Name: _____

5. Impurity measures (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 must 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.

   The attribute with the highest gain ratio is the one we will choose to split on first (this will be the first node in our tree).

SI(Passed All Assignments) $= \left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) + \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right) =$ 0.985

GR(—) $= \frac{0.004}{0.985} =$ 0.014

SI(Language) $= \left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) + \left(\frac{2}{7}\right)\log_2\left(\frac{2}{7}\right) + \left(\frac{2}{7}\right)\log_2\left(\frac{2}{7}\right) =$ 1.56

GR(—) $= \frac{0.158}{1.56} =$ 0.101

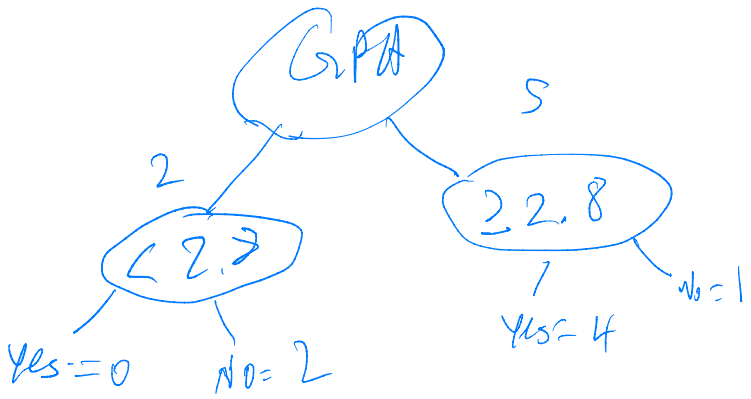
SI(GPA) $= \left(\frac{2}{7}\right)\log_2\left(\frac{2}{7}\right) + \left(\frac{5}{7}\right)\log_2\left(\frac{5}{7}\right) =$ 0.863

GR(—) $= \frac{0.261}{0.863} =$ 0.303 ← Best!

6. After you make the first split in the tree, the data records get divided up and go down different branches of the tree. The process to find the best attribute to split on is repeated for each branch of the tree.

   Calculate the Gini of the parent for all branches of the current tree.

Splitting on GPA first



Gini(< 2.7) = 0

Gini(≥ 2.8) $= 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right)$

$= 1 - 0.68$

$=$ 0.32