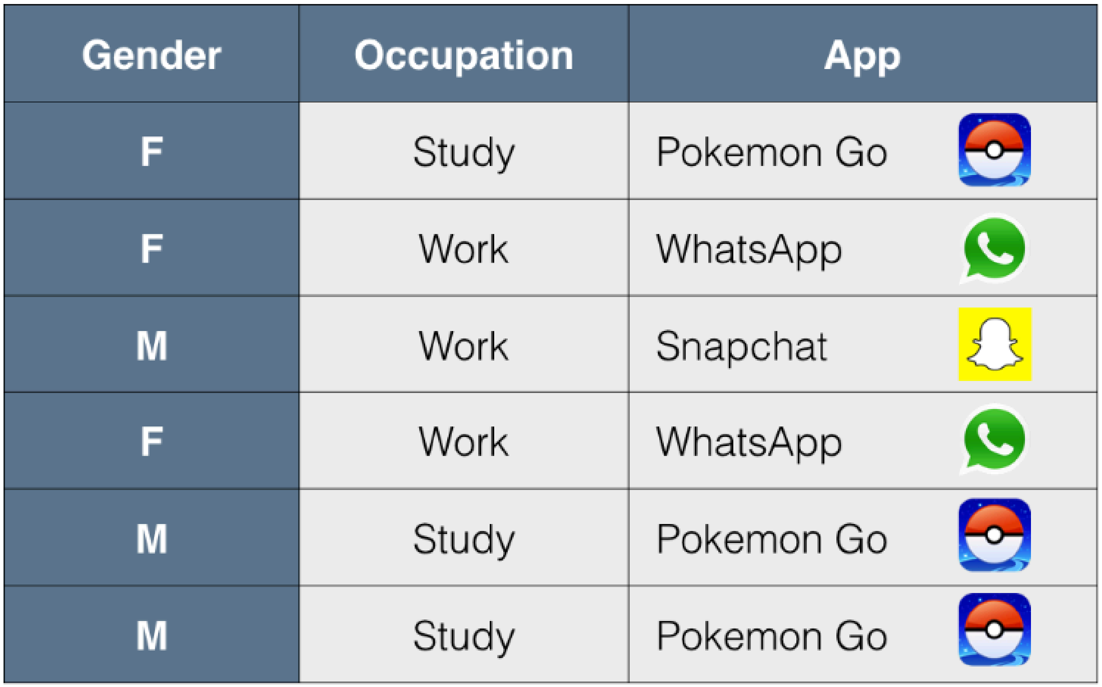


Gender	Occupation	App
F	Study	Pokemon Go
F	Work	WhatsApp
M	Work	Snapchat
F	Work	WhatsApp
M	Study	Pokemon Go
M	Study	Pokemon Go

Quiz Question

For a woman who works at an office, which app do we recommend?

a. Pokemon Go
b. WhatsApp
c. Snapchat

Quiz Question

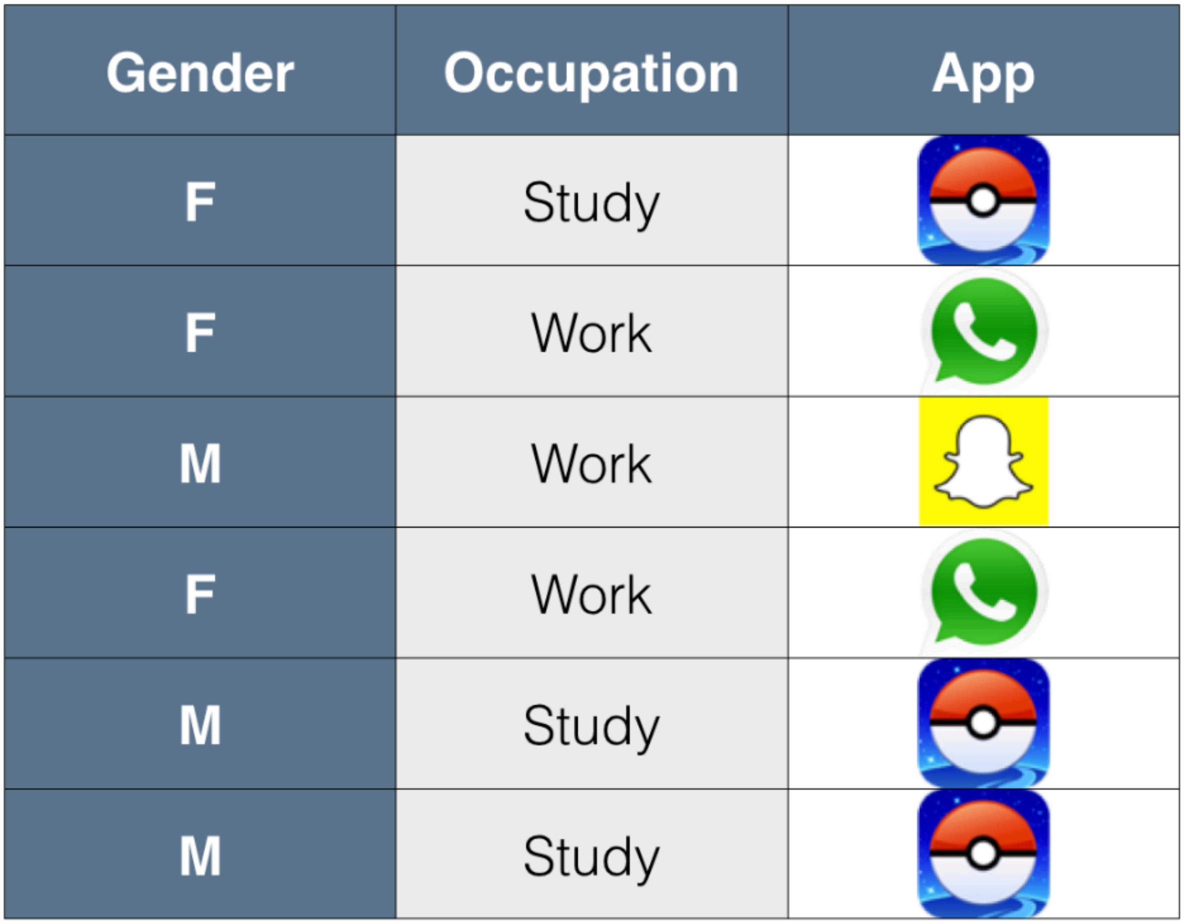
For a man who works at a factory, which app do we recommend?

a. Pokemon Go
b. Whatsapp
c. Snapchat

Quiz Question

For a girl who goes to high school, which app do we recommend?

a. Pokemon Go
b. Whatsapp
c. Snapchat

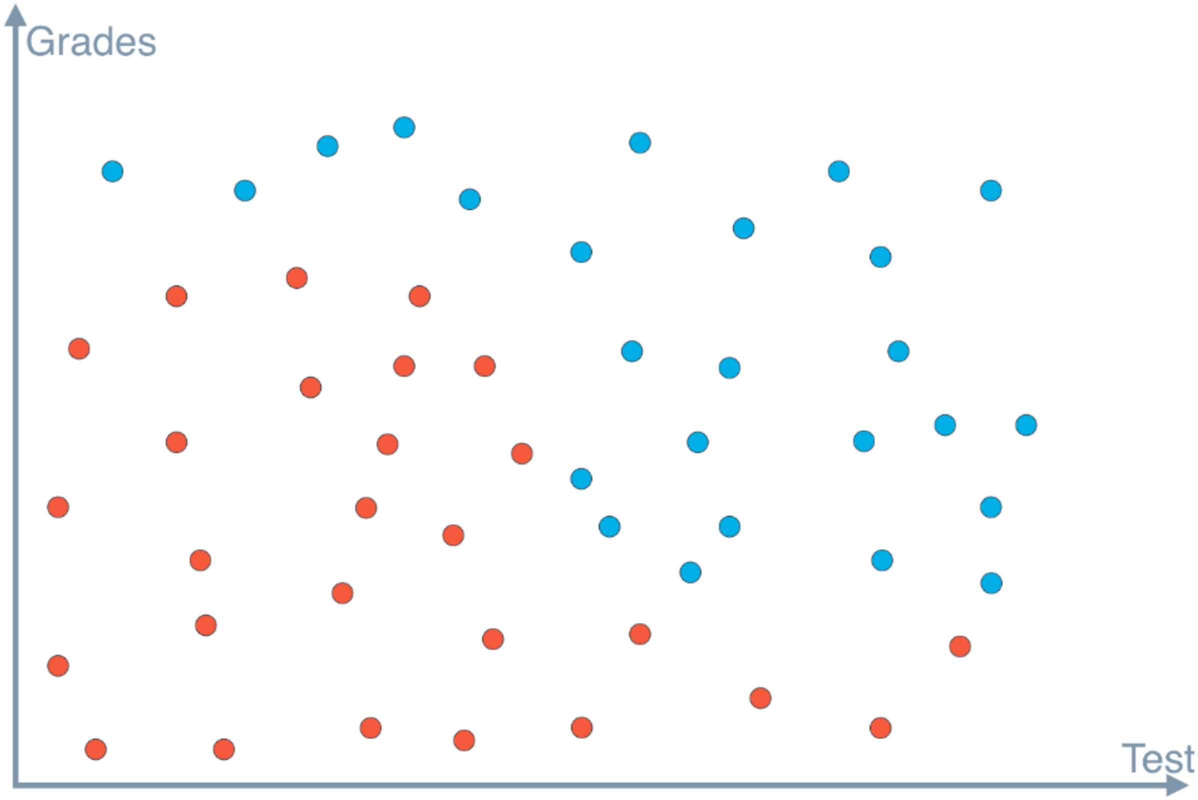


Gender	Occupation	App
F	Study	◓
F	Work	📞
M	Work	👻
F	Work	📞
M	Study	◓
M	Study	◓

Quiz Question

Between Gender and Occupation, which one seems more decisive for predicting what app will the users download?

a. Gender
b. Occupation



Quiz Question

Between a horizontal and a vertical line, which one would cut the data better?

a. Horizontal
[✓ b. Vertical]

Quiz Question

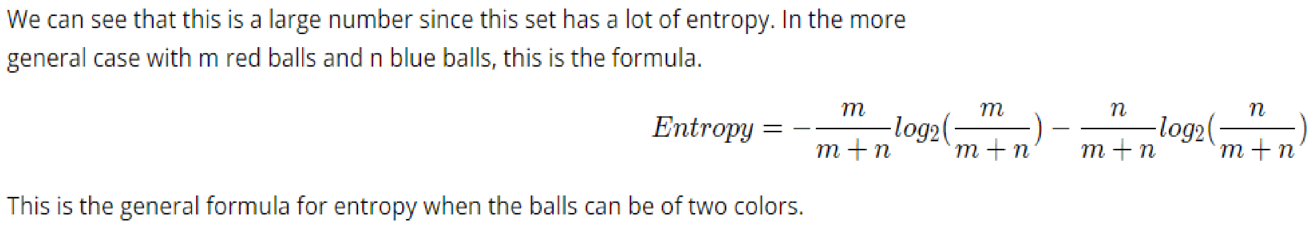
Which of the following is not accurate about entropy?

a. The more knowledge one has, the less entropy
[✓ b. The less knowledge one has, the less entropy]
c. The less knowledge one has, the more entropy

Quiz Question

Which function will help us turn products into sums?

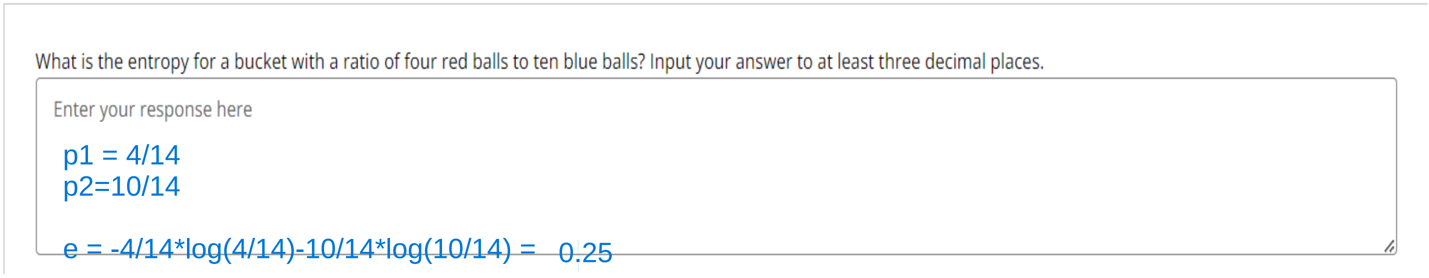
a. sin
b. cos
[✓ c. log]
d. exp



We can see that this is a large number since this set has a lot of entropy. In the more general case with m red balls and n blue balls, this is the formula.

$$Entropy = -\frac{m}{m+n}log_2(\frac{m}{m+n}) - \frac{n}{m+n}log_2(\frac{n}{m+n})$$

This is the general formula for entropy when the balls can be of two colors.



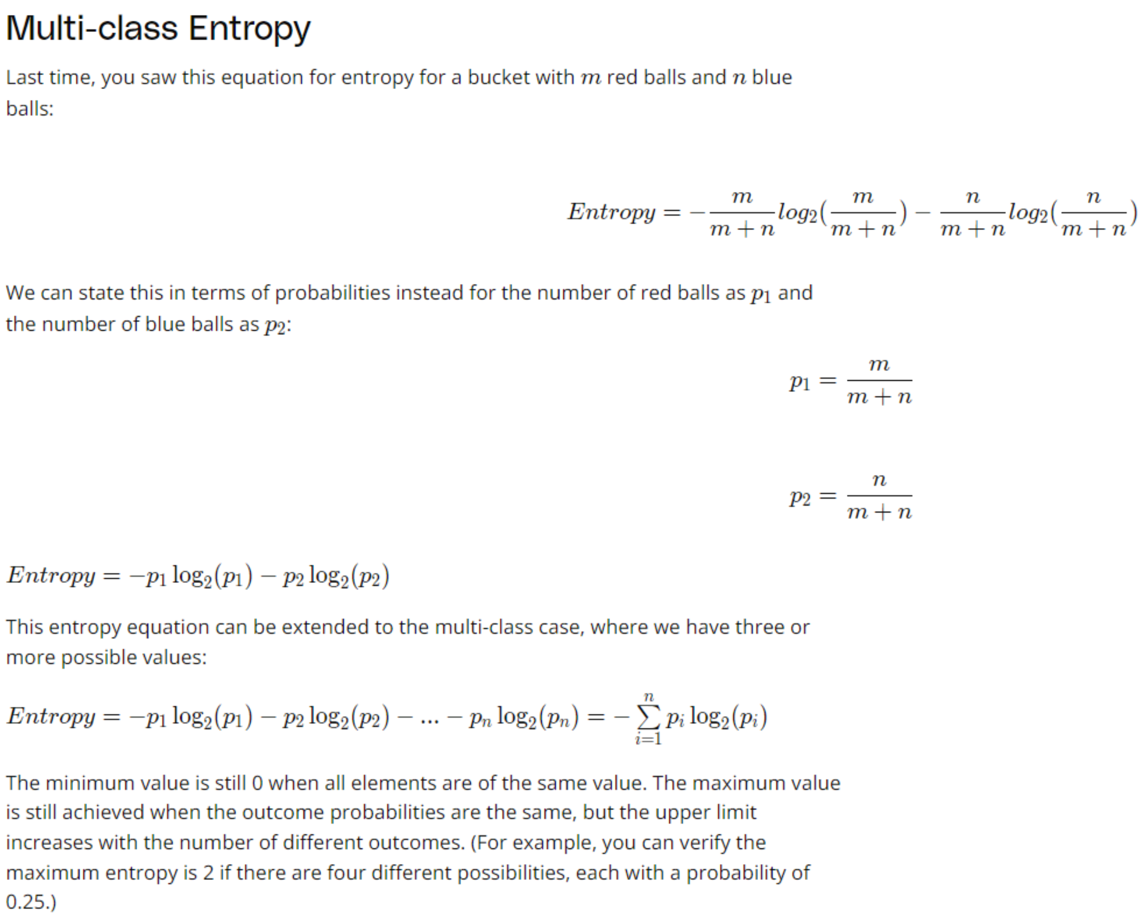
What is the entropy for a bucket with a ratio of four red balls to ten blue balls? Input your answer to at least three decimal places.

Enter your response here

p1 = 4/14
p2=10/14

e = -4/14*log(4/14)-10/14*log(10/14) = 0.25

Answer: _ _ _ _ _ _ _ _ _ _ _ _



Multi-class Entropy

Last time, you saw this equation for entropy for a bucket with m red balls and n blue balls:

$$Entropy = -\frac{m}{m+n} log_2(\frac{m}{m+n}) - \frac{n}{m+n} log_2(\frac{n}{m+n})$$

We can state this in terms of probabilities instead for the number of red balls as p_1 and the number of blue balls as p_2:

$$p_1 = \frac{m}{m+n}$$

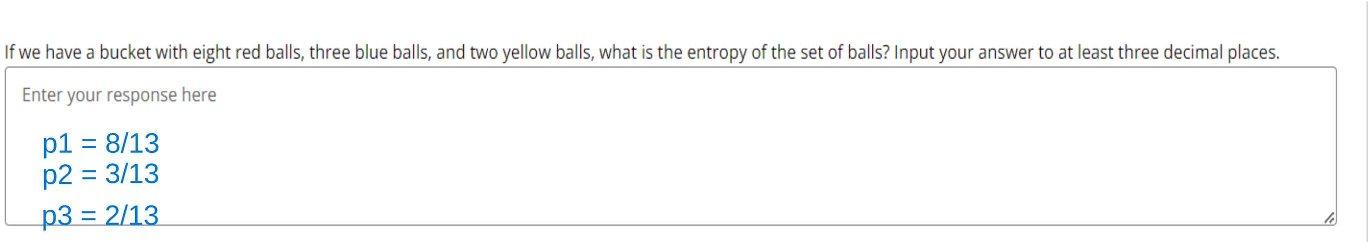
$$p_2 = \frac{n}{m+n}$$

$$Entropy = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$$

This entropy equation can be extended to the multi-class case, where we have three or more possible values:

$$Entropy = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - \ldots - p_n \log_2(p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

The minimum value is still 0 when all elements are of the same value. The maximum value is still achieved when the outcome probabilities are the same, but the upper limit increases with the number of different outcomes. (For example, you can verify the maximum entropy is 2 if there are four different possibilities, each with a probability of 0.25.)



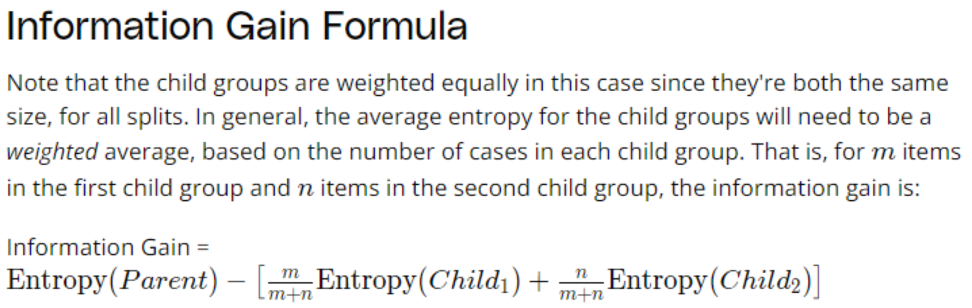
If we have a bucket with eight red balls, three blue balls, and two yellow balls, what is the entropy of the set of balls? Input your answer to at least three decimal places.

Enter your response here

p1 = 8/13
p2 = 3/13

p3 = 2/13

Answer: __ e = -8/13*log(-8/13) - 3/13*log(3/13)-2/13*log(2/13) = 0.15



Information Gain Formula

Note that the child groups are weighted equally in this case since they're both the same size, for all splits. In general, the average entropy for the child groups will need to be a *weighted* average, based on the number of cases in each child group. That is, for m items in the first child group and n items in the second child group, the information gain is:

Information Gain =

$$\mathrm{Entropy}(Parent) - \left[\frac{m}{m+n}\mathrm{Entropy}(Child_1) + \frac{n}{m+n}\mathrm{Entropy}(Child_2)\right]$$

Quiz Question

Assume we have a dataset with columns Car_Price and Car_Engine_Size and we want to recommend a specific car that is a good recommendation for the 25 age group. You calculate entropy and find that

Splitting by the Car_Price column gave us an information gain of 0.74
Splitting by the Car_Engine_Size column gave us an information gain of 0.61.

Which column will you split by?

a. [✓ Car_Price]
b. Car_Engine_Size