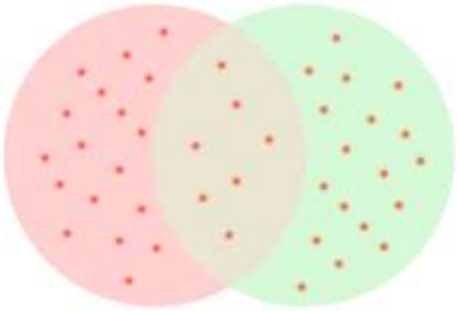
E-portfolio activity: Jaccard Coefficient Calculations

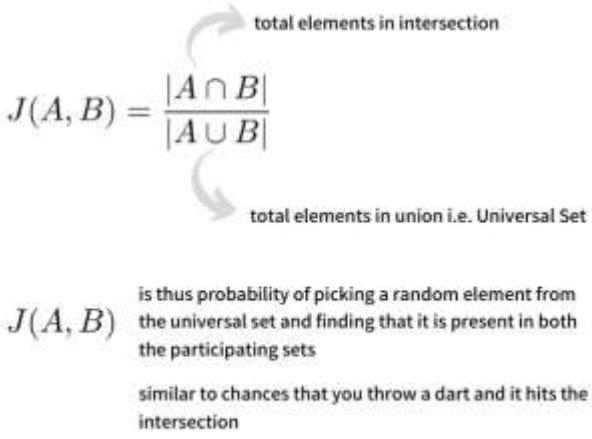
The table shows the pathological test results for three individuals.

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	A
Mary	F	Y	N	P	A	P	N
Jim	M	Y	P	N	N	N	A

Calculate Jaccard coefficient for the following pairs:

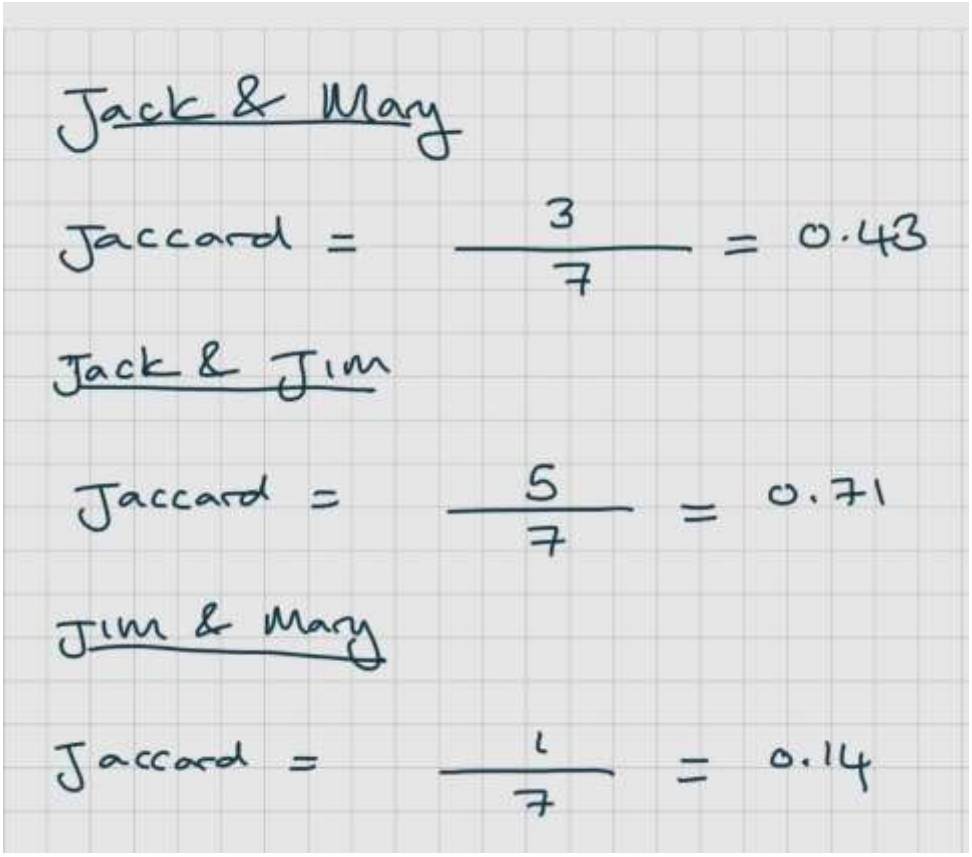
- (Jack, Mary)
- (Jack, Jim)
- (Jim, Mary)





Jaccard Similarity Coefficient as Probability

Jaccard similarity is the ratio of overlapping items to total unique items. Here are my calculations.



From this, I can conclude that Jack and Jim have the highest similarity rating in their attributes (71%), with Jim and Mary having the lowest similarity (14%).

- The Jaccard coefficient helps in tasks like clustering, classification, or identifying relationships between entities.
- In a medical context, higher similarity (e.g., Jack and Jim) might suggest these individuals have more comparable health conditions or test results. Conversely, lower similarity (e.g., Jim and Mary) indicates distinct differences that might warrant different diagnoses or treatments.

After checking the answers of this I realised my approach and calculations were wrong.

ANSWER:

To calculate the Jaccard coefficient, we first convert the asymmetric variables to binary values and re-write the table. Since Gender is a symmetric variable (that is, male, female have the same weight), it is not converted.

So let ***Y & P*** = 1; ***N & A*** = 0

$$Jaccard = \frac{f_{01}+f_{10}}{f_{01}+f_{10}+f_{11}}$$

Your final answers should be:

- (Jack, Mary) = 0.33
- (Jack, Jim) = 0.67
- (Jim, Mary) = 0.75

I attempted this again, this time creating a Jaccard function in python with the following code.

```
import numpy as np

def jaccard_binary(x,y):
    """A function for finding the similarity between two binary vectors"""
    intersection = np.logical_and(x, y)
    union = np.logical_or(x, y)
    similarity = intersection.sum() / float(union.sum())
    return similarity

# Define some binary vectors
jack = [1,0,1,0,0,0]
mary = [1,0,1,0,1,0]
jim = [1,1,0,0,0,0]

# Find similarity among the vectors
sim1 = jaccard_binary(jack,mary)
sim2 = jaccard_binary(jack,jim)
sim3 = jaccard_binary(jim,mary)

print(' Similarity between Jack and Mary is', sim1, '\n Similarity between Jack and Jim is ', sim2, '\n Similarity between Jim and Mary is ', sim3)
```

The results are as follows:

Similarity between Jack and Mary is 0.6666666666666666

Similarity between Jack and Jim is 0.3333333333333333

Similarity between Jim and Mary is 0.25

This answer is still different from the provided answer.