TASK:

*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)*

The Jaccard coefficient is a measure of similarity between data sets. Higher scores indicate more similar populations.

To start, we convert current values into numeric for calculations. For symptoms and test results, we convert yes/present /positive to 1, no/absent/negative to 0. This assigns the non-zero value for all “true” conditions. Since Jaccard is for “present” features, we ignore the Gender column.

| ***Name*** | ***Gender*** | ***Fever*** | ***Cough*** | ***Test-1*** | ***Test-2*** | ***Test-3*** | ***Test-4*** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Jack* | *M* | *1* | *0* | *1* | *0* | *0* | *0* |
| *Mary* | *F* | *1* | *0* | *1* | *0* | *1* | *0* |
| *Jim* | *M* | *1* | *1* | *0* | *0* | *0* | *0* |

Present:

Jack: Fever, Test 1 – [1,0,1, 0, 0, 0]

Mary: Fever, Test 1, Test 3 – [1,0,1, 0, 1, 0]

Jim: Fever, Cough – [1,1, 0, 0, 0, 0]

Jaccard distance is calculated as:

f01 + f10

f01 + f10 + f11

where for pair (A,B):

f01: A=0, B=1

f10: A=1, B=0

f11: A=1, B=1

Calculation: Jack ([1,0,1, 0, 0, 0]), Mary ([1,0,1, 0, 1, 0])

1 + 0

1 + 0 + 2

(Jack, Mary) = 0.33

Calculation: Jack ([1,0,1, 0, 0, 0]), Jim ([1,1, 0, 0, 0, 0])

1 + 1

1 + 1 + 1

(Jack, Jim) = 0.67

Calculation: Mary ([1,0,1, 0, 1, 0]), Jim ([1,1, 0, 0, 0, 0])

1 + 2

1 + 2 + 1

(Mary, Jim) = 0.75