Unit 5: Clustering
Unit 5 Artefacts

Wiki Activity: Clustering

I watched this animation and read the blog at:
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

This helped me relate the algorithm logic to the outputs in greater detail. I find learning models visually is very effective for me. I found it very helpful to see how choosing the initial starting points of the centroids can affect the performance and output. The animation made that easier to understand. However, one consideration it did give me was the understanding of social and ethical issues, because initialisation parameters are chosen by the analyst, this introduces the possibility of bias or even discrimination. Manipulation of parameters like these can be abused to obtain results that are more valuable to certain groups and may be detrimental to other groups. This can also have legal and professional implications and machine learning professionals should try and overcome these challenges by being open, transparent and mindful of how their methods could be affected by conscious and unconscious biases alike.

e-Portfolio Activity: Jaccard Coefficient Calculations

Jaccard coefficients are a statistic used for understanding the similarity or diversity in a sample set. It is essentially defined as the size of the intersection divided by the size of the union of the sample sets.

First, I converted yes/positive into a 1 or no/absent into 0:

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

In the context of binary variables, the coefficient is calculated using the following formula available in the ML_PCOM7E March 2024 module content information:

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

f01 is the number of attributes where the first set has a value of 0 and the second set has a value of 1.
f10 is the number of attributes where the first set has a value of 1 and the second set has a value of 0.
f11 is the number of attributes where both sets have a value of 1.

Then, I used the equation above as per the

For Jack and Mary:

To compute the Jaccard coefficient, we need the following values:

f01: Number of attributes where Jack has 0 and Mary has 1; or 1 (Test-3)
f10: Number of attributes where Jack has 1 and Mary has 0; or 0
f11: Number of attributes where both have 1; or 2 (Fever and Test-1)

Therefore: Jaccard(Jack, Mary) = 0.33

I then repeated this for (Jack, Jim):
f01: 0
f10: 2 (Cough and Test-1)
f11: 1 (Fever)

Jaccard(Jack, Jim) = 0.67

And (Jim, Mary):

f01: 2 (Cough and Test-3)
f10: 1 (Test-1)
f11: 1 (Fever)

Jaccard(Jim, Mary) = 0.75