UNIT 5

Portfolio Activity: Jaccard Coefficient Calculations

Murthy Kanuri
Machine Learning
University of Essex

Table of Contents

1 Scenario

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	M	N	N	N

Calculate Jaccard Coefficient for the following pairs:

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

2 What is Jaccard Coefficient

- A Commonly used measure of overlap of two sets A and B is the Jaccard coefficient
- Jaccard (A, B) = $\frac{|A \cap B|}{|A \cup B|}$ or Jaccard (J) = $\frac{f_{01} + f_{10}}{(f_{01} \ + f_{10} \ + f_{11})}$

3 Calculate Jaccard Coefficient for (Jack, Mary)

Converting asymmetric variables into binary format and re-write the table

- Fever: N (0), Y (1)
- Cough: N (0), P (1)
- Test: N or A (0), P (1)

Attribute	Jack	Mary	Observation
Fever	1	1	f_{11} (1,1)
Cough	0	0	Ignore as both 0
Test-1	1	1	f_{11} (1,1)
Test-2	0	0	Ignore as both 0
Test-3	0	1	f_{01} (0.1)
Test-4	0	0	Ignore as both 0

From the above table

- f_{11} (1,1) = Fever + Test-1 = 2
- f_{01} (0.1) = Test-3 = 1
- f_{10} = 0

Jaccard Coefficient for (Jack, Mary) =$\frac{f_{01}+f_{10}}{(f_{01}\ +f_{10}\ +f_{11)}}$ = $\frac{1+0}{1+0+2}$ = $\frac{1}{3}$ = 0.33

4 Calculate Jaccard Coefficient for (Jack, Jim)

Converting asymmetric variables into binary format and re-write the table

- Fever: N (0), Y (1)
- Cough: N (0), P (1)
- Test: N or A (0), P (1)

Attribute	Jack	Jim	Observation
Fever	1	1	f_{11} (1,1)
Cough	0	1	f_{01} (0,1)
Test-1	1	0	f_{10} (1,0)
Test-2	0	0	Ignore as both 0
Test-3	0	0	Ignore as both 0
Test-4	0	0	Ignore as both 0

From the above table

- f_{11} (1,1) = Fever = 1
- f_{01} (0.1) = Cough = 1
- f_{10} (1,0) = Test -1 = 1

Jaccard Coefficient for (Jack, Mary) =$\frac{f_{01}+f_{10}}{(f_{01}\ +f_{10}\ +f_{11)}}$ $\frac{1+1}{1+1+1}$ = $\frac{2}{3}$ = 0.67

5 Calculate Jaccard Coefficient for (Jim, Mary)

Converting asymmetric variables into binary format and re-write the table

- Fever: N (0), Y (1)
- Cough: N (0), P (1)
- Test: N or A (0), P (1)

Attribute	Jim	Mary	Observation

Fever	1	1	f_{11} (1,1)
Cough	1	0	f_{10} (1,0)
Test-1	0	1	f_{01} (0.1)
Test-2	0	0	Ignore as both 0
Test-3	0	1	f_{01} (0.1)
Test-4	0	0	Ignore as both 0

From the above table

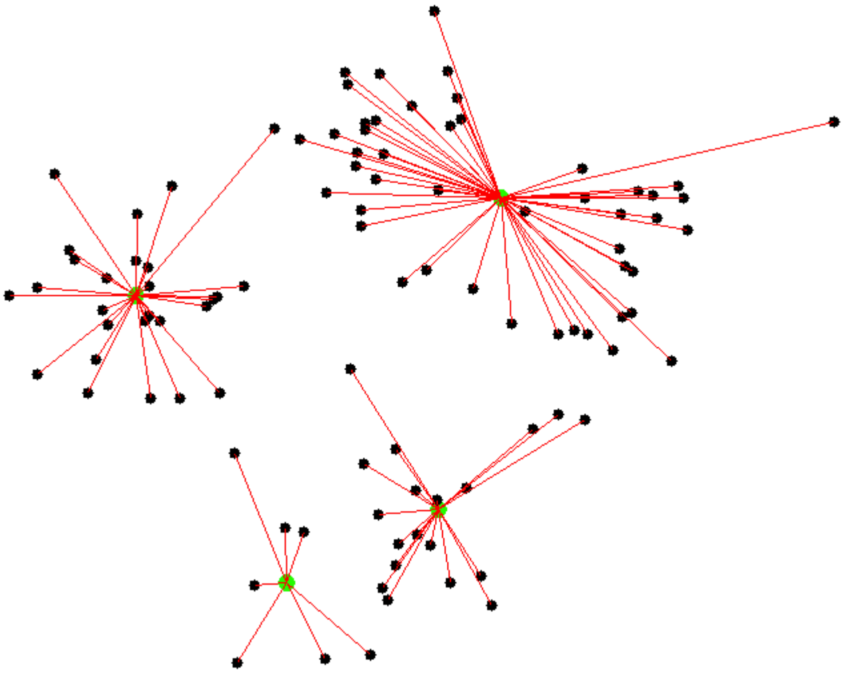
- f_{11} (1,1) = Fever = 1
- f_{01} (0.1) = Test-1 + Test-3 = 2
- f_{10} = 0

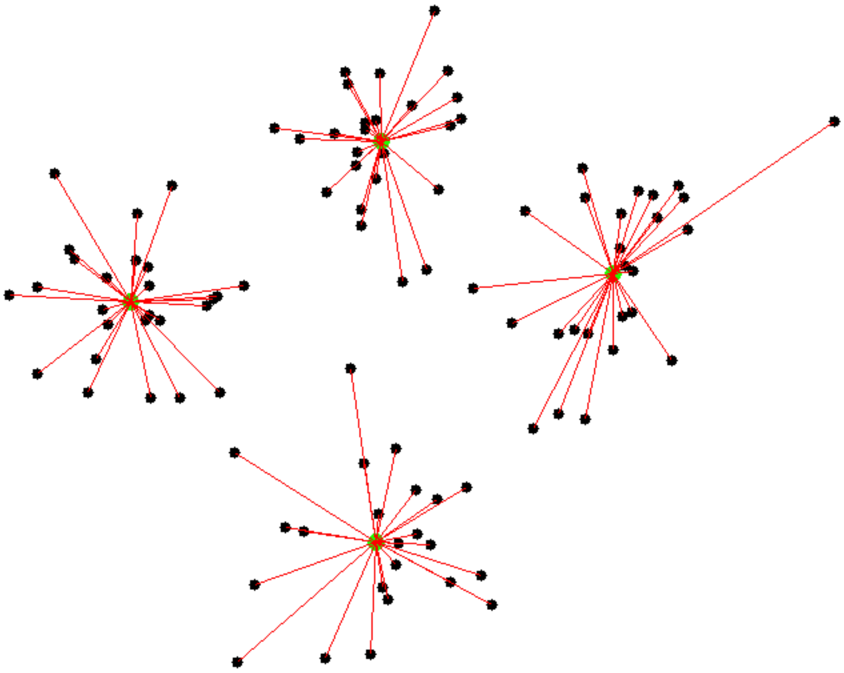
Jaccard Coefficient for (Jack, Mary) = $\frac{f_{01}+f_{10}}{(f_{01}+f_{10}+f_{11})} = \frac{2+0}{2+0+1} = \frac{2}{3}$ = 0.75

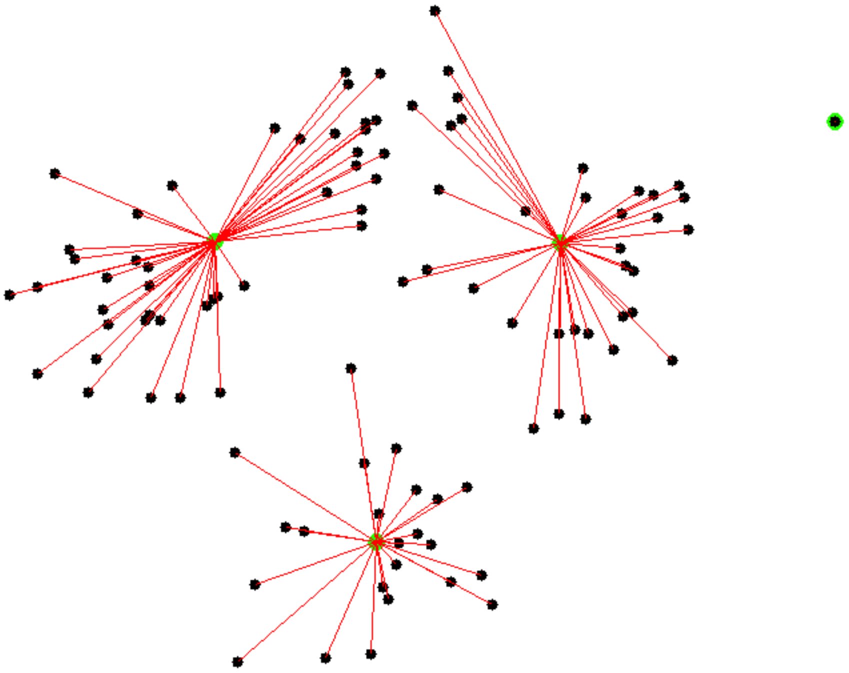
5.1 Legal, Social, Ethical, and Professional Considerations in Machine Learning

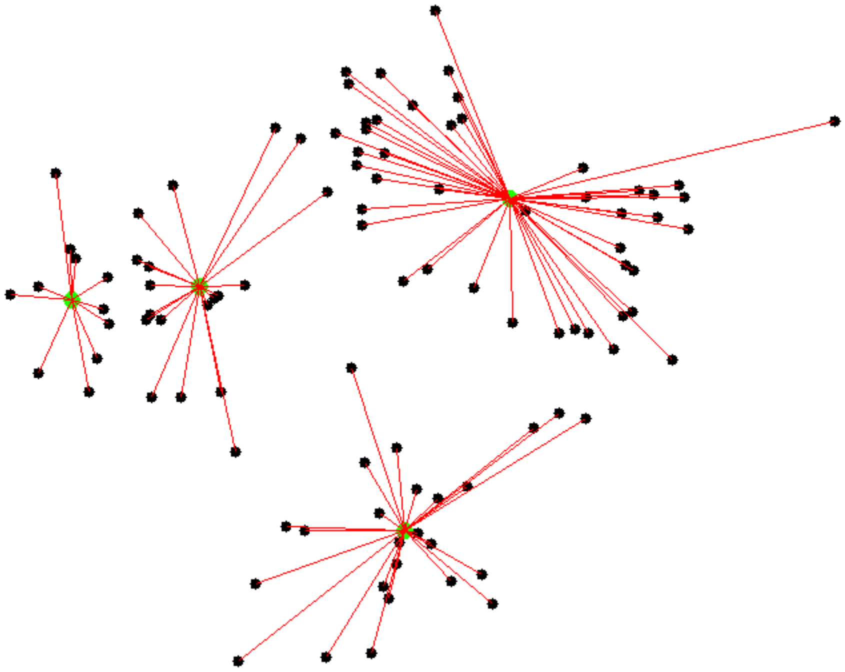
- Ensuring GDPR Compliance or HIPAA in Healthcare clustering
- Data biases can lead to discrimination or unethical profiling
- All the stakeholders should have transparency about the data being used
- Qualified professionals should be made responsible for algorithmic decisions to follow ethical standards
- Ensure that datasets used for clustering are accurate and true representatives.

5.2 Images for Reference (Shabal.in)

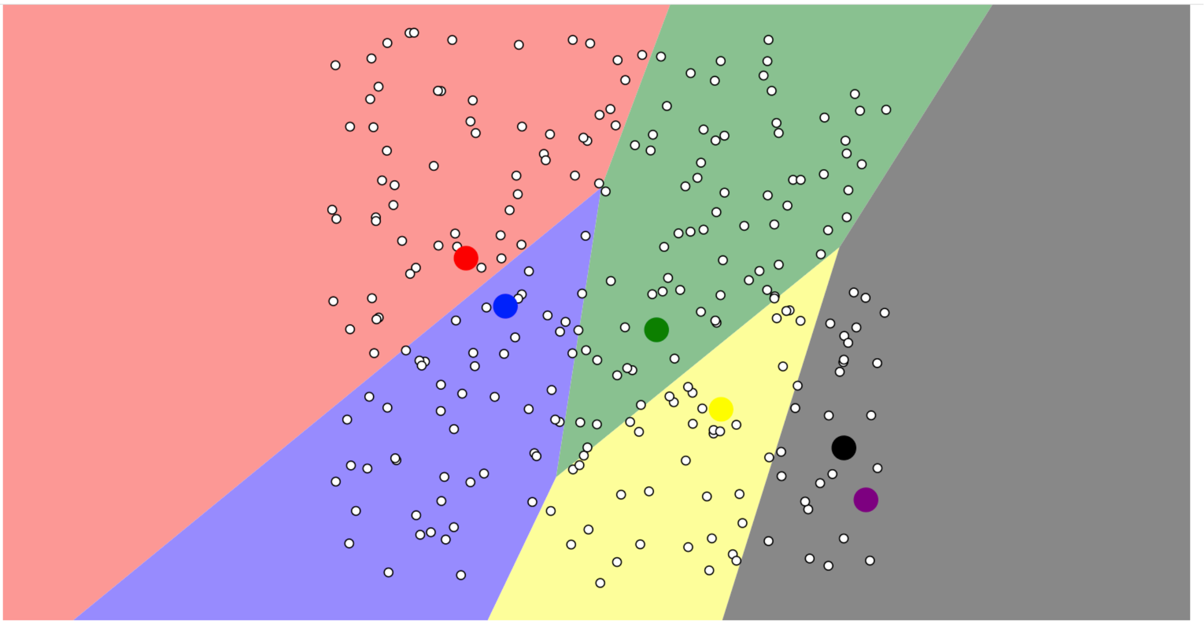


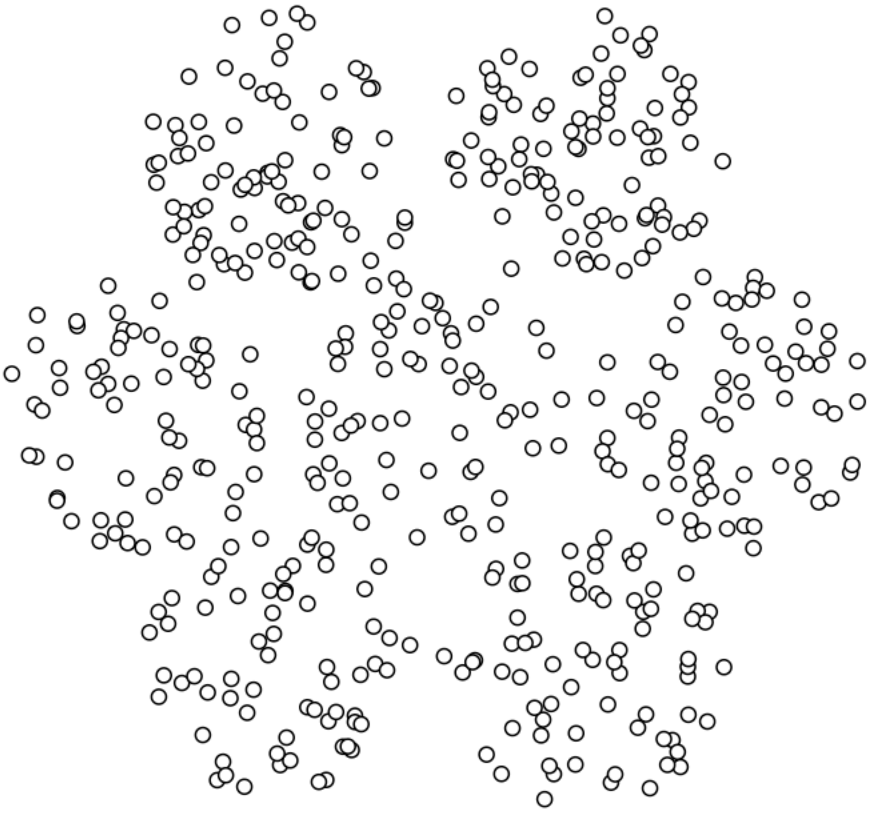






5.3 Images for Reference (Naftali Harris)





References

Shabal. (n.d.) K-Means Clustering Visualization. Available at: https://shabal.in/visuals/kmeans/2.html (Accessed: 10 November 2024).

Harris, N. (n.d.) Visualizing K-Means Clustering. Available at: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/ (Accessed: 10 November 2024).

Patel, S. (2019) K-means Clustering Algorithm: Implementation and Critical Analysis. Saarbrücken: Scholars' Press.

Dwivedi, G. (2023) Optimization of K-Means Clustering Using Genetic Algorithm. Saarbrücken: LAP LAMBERT Academic Publishing.