Instituto Politécnico Nacional
Escuela Superior de Cómputo
Unidad de Aprendizaje: Minería de datos
Ciclo escolar: 2022–2

Proyecto No. 4. Clúster

Grupo: 5CDM1 Equipo: 1

Nombre de los integrantes del equipo:

1) De Luna Ocampo Yanina

2) Medina Barreras Daniel Iván

INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

LICENCIATURA EN CIENCIA DE DATOS

UNIDAD DE APRENDIZAJE

MINERÍA DE DATOS

CREACIÓN DE CLUSTERS - MATRIZ

NOMBRE DE LOS ALUMNOS:

DE LUNA OCAMPO YANINA

MEDINA BARRERAS DANIEL IVÁN

PROFESOR:

OCAMPO BOTELLO FABIOLA

GRUPO:

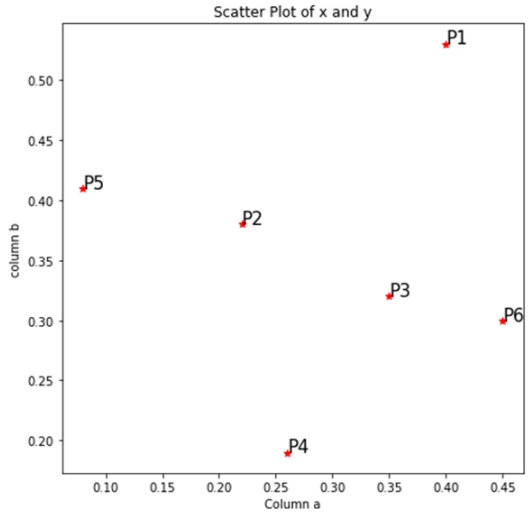
5CDM1

FECHA:

29/05/2022

PARTE 1. Diseñar una matriz con 6 puntuaciones (x, y) y desarrollar el proceso paso a paso de la creación de grupos aplicando las tres siguientes técnicas (*Inserte una portada indicando el tema que aborda)*

Point	a	b
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30



INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

LICENCIATURA EN CIENCIA DE DATOS

UNIDAD DE APRENDIZAJE

MINERÍA DE DATOS

CREACIÓN DE CLUSTERS - ENLACE SIMPLE

NOMBRE DE LOS ALUMNOS:

DE LUNA OCAMPO YANINA

MEDINA BARRERAS DANIEL IVÁN

PROFESOR:

OCAMPO BOTELLO FABIOLA

GRUPO:

5CDM1

FECHA:

29/05/2022

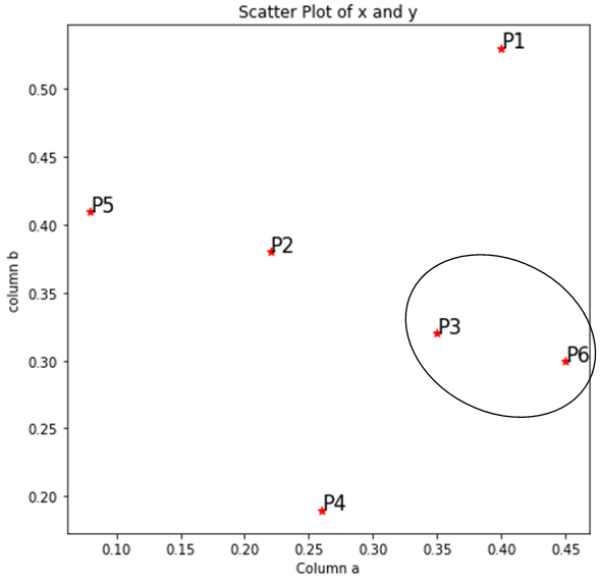
1. Enlace simple

Matriz de distancias

Primer valor mínimo

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

Gráfico primer grupo



Actualizar matriz de distancias después crear primer grupo

MIN(DIST((P3,P1),(P6,P1))) = MIN[(0.22,0.23)] = 0.22

MIN(DIST((P3,P2),(P6,P2))) = MIN[(0.15,0.25)] = 0.15

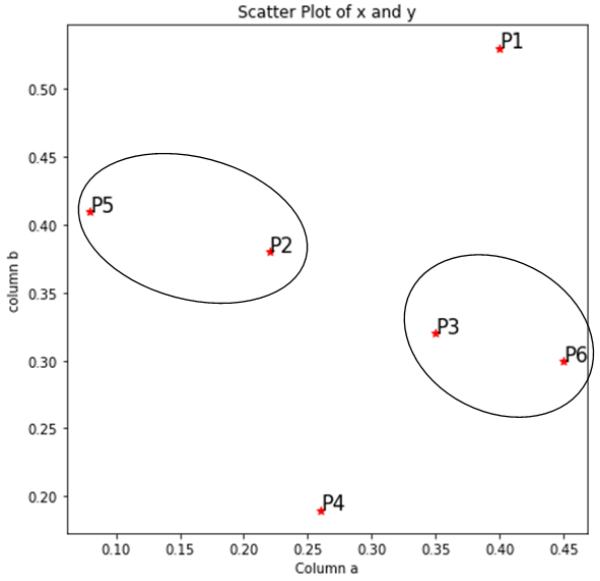
MIN(DIST((P3,P4),(P6,P4))) = MIN[(0.15,0.22)] = 0.15

MIN(DIST((P3,P5),(P6,P5))) = MIN[(0.28,0.39)] = 0.22

Matriz actualizada para P3,P6 y nuevo valor mínimo

	P1	P2	P3,P6	P4	P5
P1	0				
P2	0.24	0			
P3,P6	0.22	0.15	0		
P4	0.37	0.20	0.15	0	
P5	0.34	0.14	0.28	0.29	0

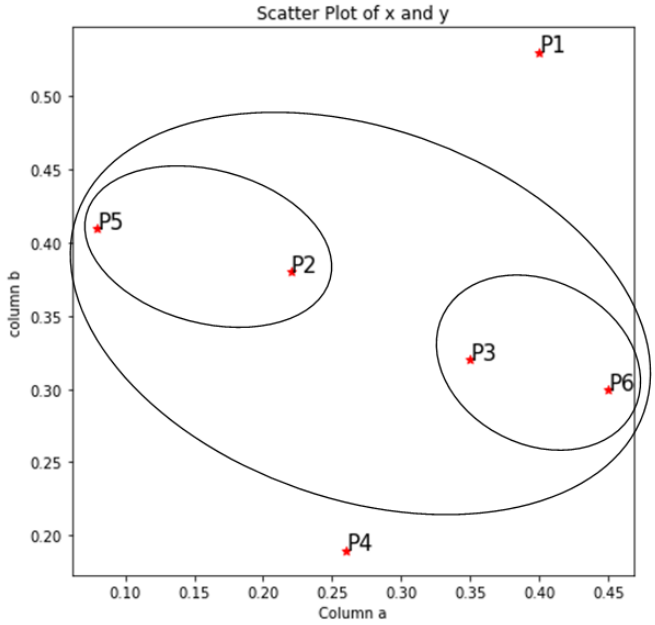
Gráfico segundo grupo



MIN(DIST((P2,P1),(P5,P1))) = MIN[(0.23,0.34)] = 0.23
MIN(DIST((P2,(P3,P6)),(P5,(P3,P6)))) = MIN[(0.15,0.28)] = 0.15
MIN(DIST((P2,P4),(P5,P4))) = MIN[(0.20,0.29)] = 0.20
Matriz actualizada P5,P2

	P1	P2, P5	P3,P6	P4
P1	0			
P2, P5	0.23	0		
P3,P6	0.22	0.15	0	
P4	0.37	0.20	0.15	0

Gráfico para el tercer grupo

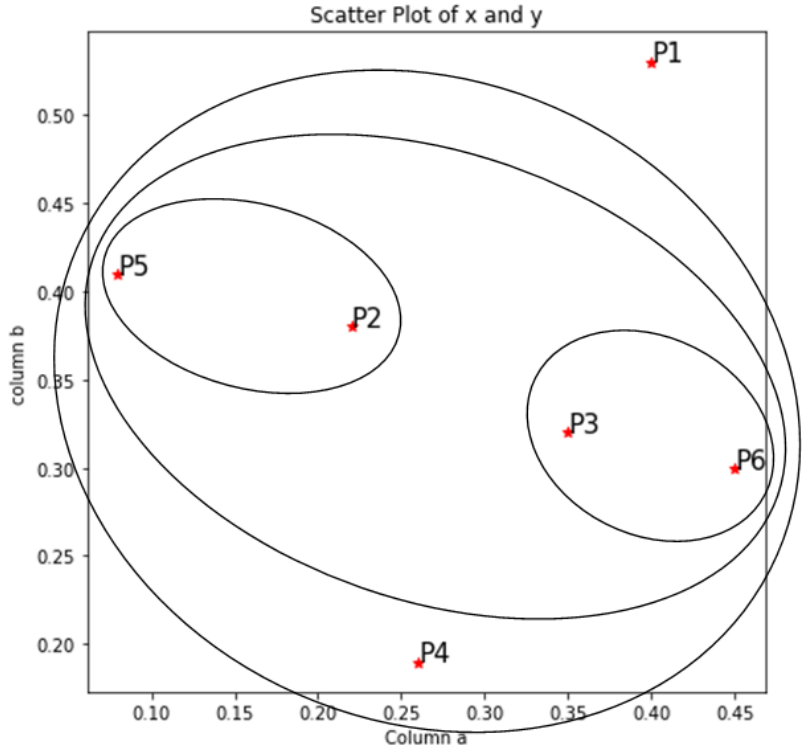


MIN(DIST(((P2, P5),P1),((P3,P6),P1))) = MIN[(0.23,0.22)] = 0.22
MIN(DIST(((P2,P5),P4),((P3,P6),P4))) = MIN[(0.20,0.15)] = 0.15

Matriz actualizada P2,P5,P3, P6

	P1	P2, P5,P3,P6	P4
P1	0		
P2, P5, P3, P6	0.22	0	
P4	0.37	0.15	0

Gráfica 4to grupo

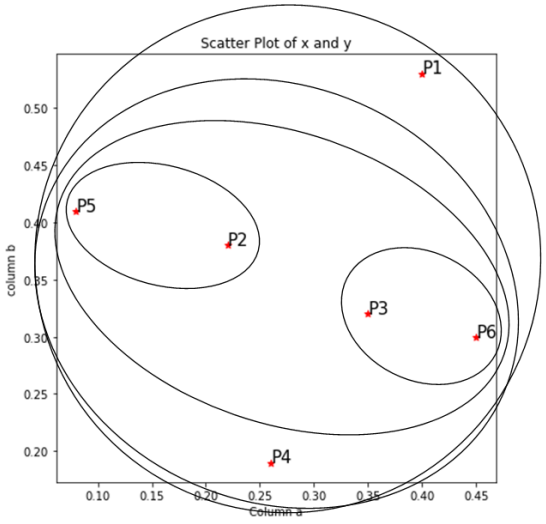


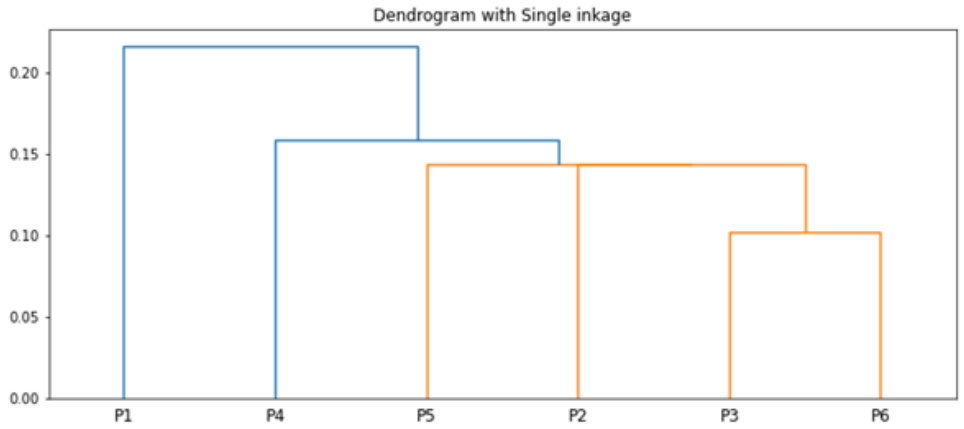
MIN(DIST((((p2,p5,p3,p6)),P1),((P4,P1))) = MIN[(0.22,0.37)] = 0.22

Matriz actualizada para P2,P5,P3,P6,P4

	P1	P2, P5,P3,P6,P4
P1	0	
P2, P5, P3, P6,P4	0.22	0

Gráfica final




Dendrogram with Single inkage
0.20
0.15
0.10
0.05
0.00
P1
P4
P5
P2
P3
P6

INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

LICENCIATURA EN CIENCIA DE DATOS

UNIDAD DE APRENDIZAJE

MINERÍA DE DATOS

CREACIÓN DE CLUSTERS - ENLACE PROMEDIO

NOMBRE DE LOS ALUMNOS:

DE LUNA OCAMPO YANINA

MEDINA BARRERAS DANIEL IVÁN

PROFESOR:

OCAMPO BOTELLO FABIOLA

GRUPO:

5CDM1

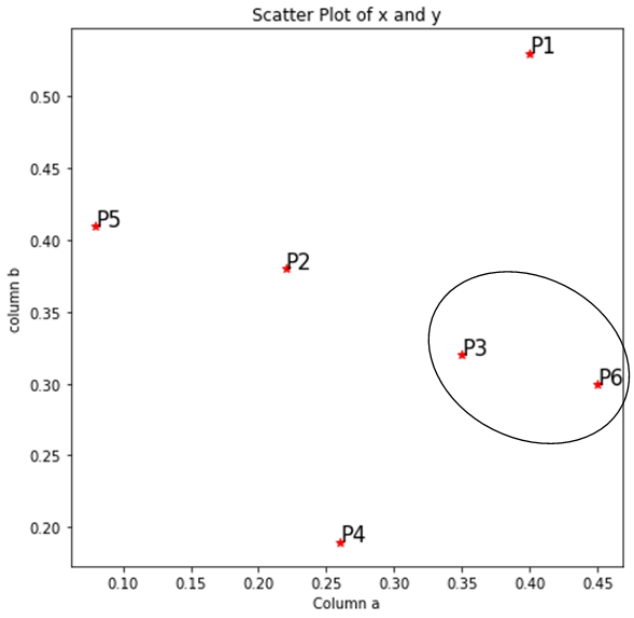
FECHA:

29/05/2022

2. Enlace promedio

Matriz de distancias
Primer valor mínimo

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0



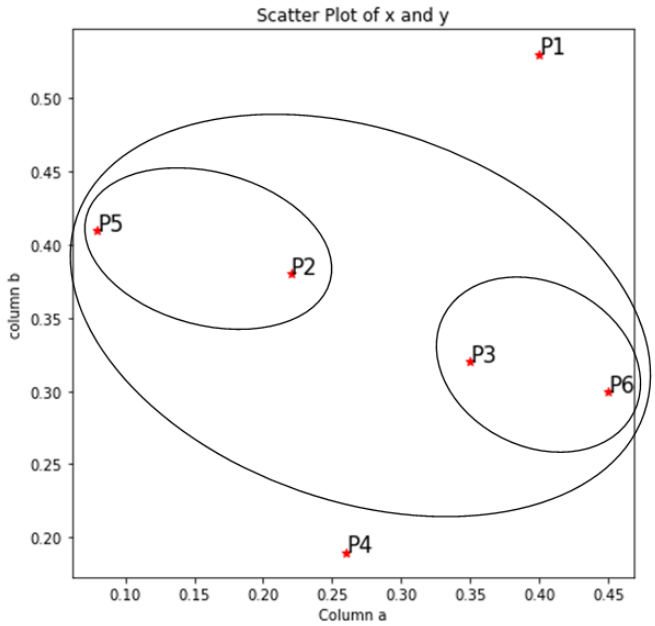
Actualizar matriz de distancias después crear primer grupo
AVG(DIST((P3,P6),P1)) =
DIST[(P3,P1),P1] = ½(DIST(P3,P1)+DIST(P6,P1)=½(0.22+0.23)=½(0.45)=0.23
AVG(DIST((P3,P6),P2)) =
DIST[(P3,P6),P2] = ½(DIST(P3,P2)+DIST(P6,P2)=½(0.15+0.25)=½(0.4)=0.2
AVG(DIST((P3,P6),P4)) =
DIST[(P3,P6),P4] = ½(DIST(P3,P4)+DIST(P6,P2)=½(0.15+0.22)=½(0.37)=0.19
AVG(DIST((P3,P6),P5)) =

DIST[(P3,P6),P5] = ½(DIST(P3,P5)+DIST(P6,P5)=½(0.28+0.39)=½(0.67)=0.34

Matriz actualizada para P3, P6, más el valor mínimo nuevo

	P1	P2	P3,P6	P4	P5
P1	0				
P2	0.24	0			
P3,P6	0.23	0.2	0		
P4	0.37	0.20	0.19	0	
P5	0.34	0.14	0.34	0.29	0

Gráfico para el tercer grupo



AVG(DIST((P2,P5),P1)) =
DIST[(P2,P5),P1] = ½(DIST(P2,P1)+DIST(P5,P1)=½(0.23+0.34)=½(0.57)=0.29

AVG(DIST((P2,P5),(P3,P6))) =
DIST[(P2,P5),(P3,P6)] =
½(DIST(P2,(P3,P6))+DIST(P5,(P3,P6))=½(0.2+0.34)=½(0.54)=0.27

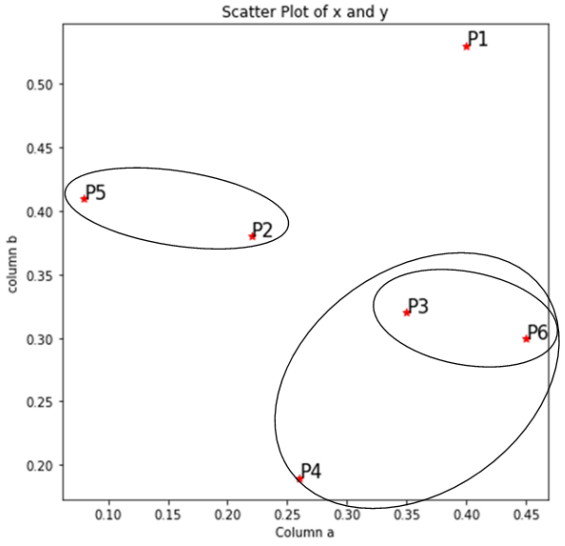
AVG(DIST((P2,P5),P4)) =

DIST[(P2,P5),P4] = ½(DIST(P2,(P4))+DIST(P5,P4))=½(0.20+0.29)=½(0.49)=0.25

Matriz actualizada P2,P5

	P1	P2, P5	P3,P6	P4
P1	0			
P2, P5	0.29	0		
P3,P6	0.22	0.27	0	
P4	0.37	0.25	0.19	0

Gráfica actualizada tercer grupo

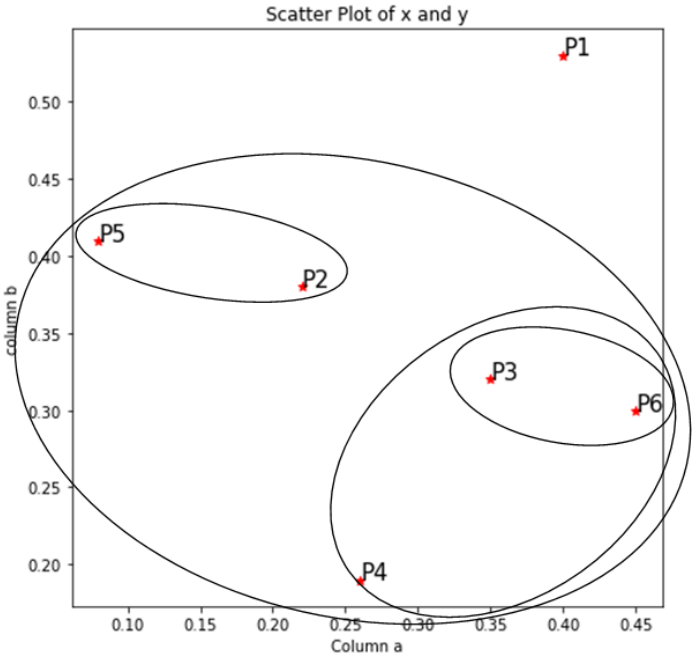


AVG(DIST((P3,P6,P4),P1)) =
DIST[(P3,P6,P4),P1] = ½(DIST((P3,P6),P1)+DIST((P4),P1)=½(0.23+0.37)=½(0.6)=0.3

AVG(DIST((P3,P6,P4),(P2,P5))) =
DIST[(P3,P6,P4),(P2,P5)] =
½(DIST((P3,P6),(P2,P5))+DIST((P4),(P2,P5))=½(0.27+0.25)=½(0.52)=0.26

	P1	P2, P5,	P3,P6,P4
P1	0		
P2, P5,	0.29	0	
P3,P6,P4	0.3	0.26	0

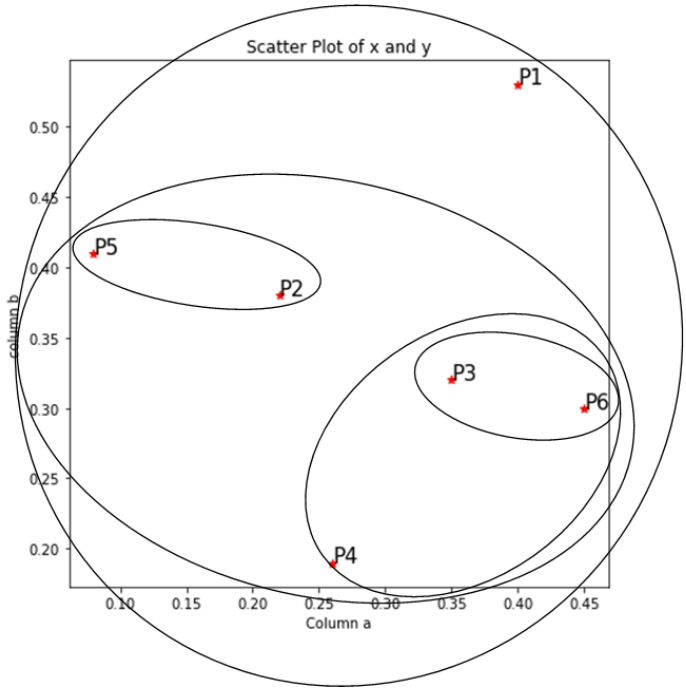
Cuarto grupo

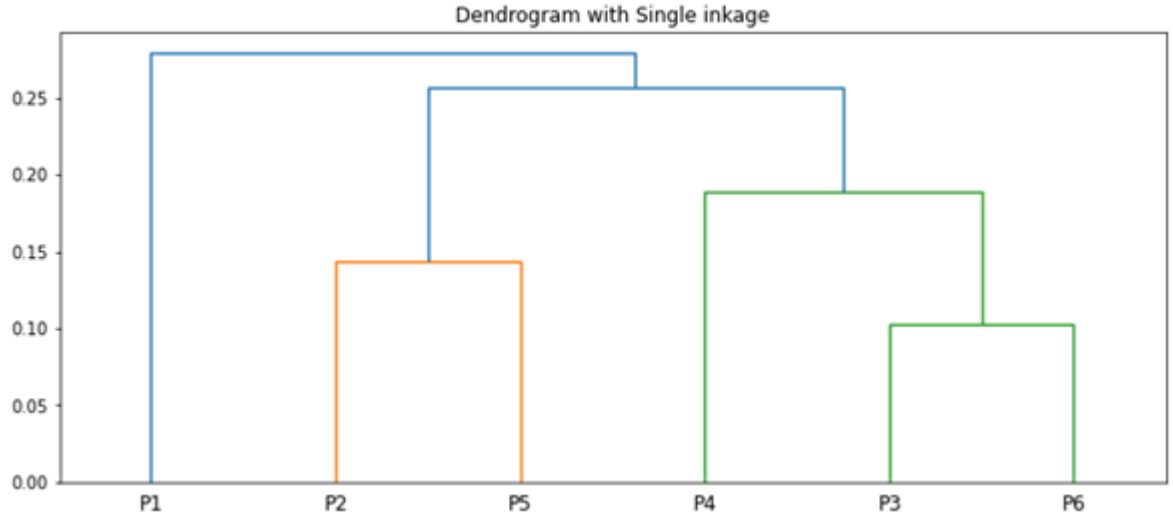


AVG(DIST((P3,P6,P4),(P2,P5))) =
DIST[(P3,P6,P4),(P2,P5)] =
½(DIST((P3,P6,P4),P1)+DIST((P2,P5),P1)=½(0.3+0.29)=½(0.59)=0.3

	P1	P2, P5,P3,P6,P4
P1	0	
P2, P5, P3, P6,P4	0.3	0

Grupo final


Scatter Plot of x and y
P1
P2
P3
P4
P5
P6
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
Column a
Column b


Dendrogram with Single inkage
0.25
0.20
0.15
0.10
0.05
0.00
P1
P2
P5
P4
P3
P6

INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

LICENCIATURA EN CIENCIA DE DATOS

UNIDAD DE APRENDIZAJE

MINERÍA DE DATOS

CREACIÓN DE CLUSTERS - ENLACE COMPLETO

NOMBRE DE LOS ALUMNOS:

DE LUNA OCAMPO YANINA

MEDINA BARRERAS DANIEL IVÁN

PROFESOR:

OCAMPO BOTELLO FABIOLA

GRUPO:

5CDM1

FECHA:

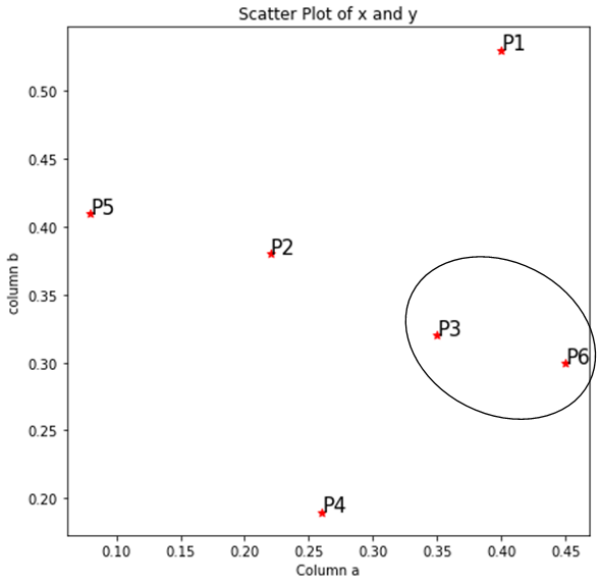
29/05/2022

3. Enlace completo

Matriz de distancias
Primer valor mínimo

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

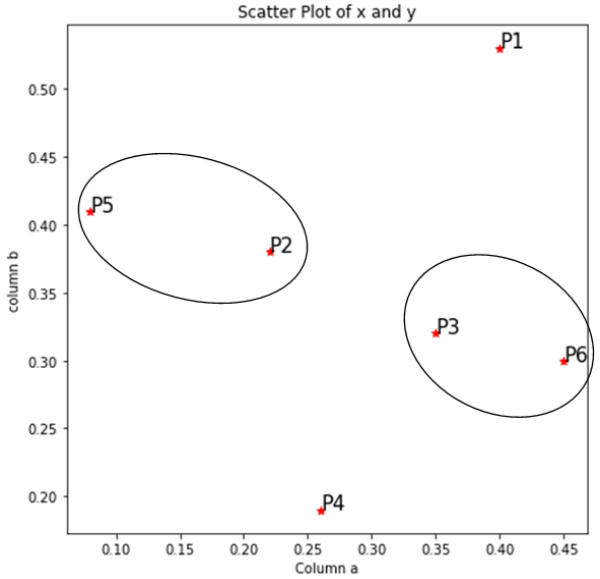
Primer grupo



Actualizar matriz de distancias después crear primer grupo
MAX(DIST((P3,P6),P1))) = MAX(DIST((P3,P1),(P6,P1)))) = MAX[(0.22,0.23)] = 0.23
MAX(DIST((P3,P6),P2))) = MAX(DIST((P3,P2),(P6,P2)))) = MAX[(0.15,0.25)] = 0.25
MAX(DIST((P3,P6),P4))) = MAX(DIST((P3,P4),(P6,P4)))) = MAX[(0.15,0.22)] = 0.22
MAX(DIST((P3,P6),P5))) = MAX(DIST((P3,P2),(P6,P5)))) = MAX[(0.28,0.39)] = 0.39

	P1	P2	P3,P6	P4	P5
P1	0				
P2	0.23	0			
P3,P6	0.23	0.25	0		
P4	0.37	0.20	0.2	0	
P5	0.34	0.14	0.39	0.29	0

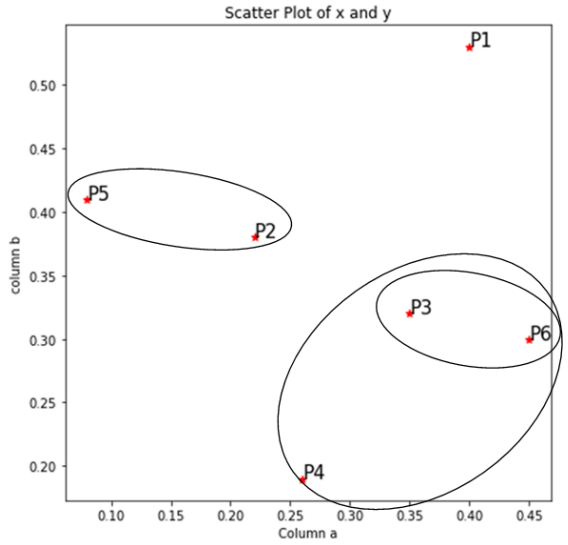
Segundo grupo



MAX(DIST((P2,P5),P1))) = MAX(DIST((P2,P1),(P5,P1)))) = MAX[(0.23,0.34)] = 0.34
MAX(DIST((P2,P5),(P3,P6)))) = MAX(DIST((P2,(P3,P6),(P5,(P3,P6)))) =
MAX[(0.25,0.39)] = 0.39
MAX(DIST((P2,P5),P4))) = MAX(DIST((P2,P4),(P5,P4)))) = MAX[(0.20,0.29)] = 0.29

	P1	P2, P5	P3,P6	P4
P1	0			
P2, P5	0.34	0		
P3,P6	0.23	0.39	0	
P4	0.37	0.29	0.22	0

Cuarto grupo

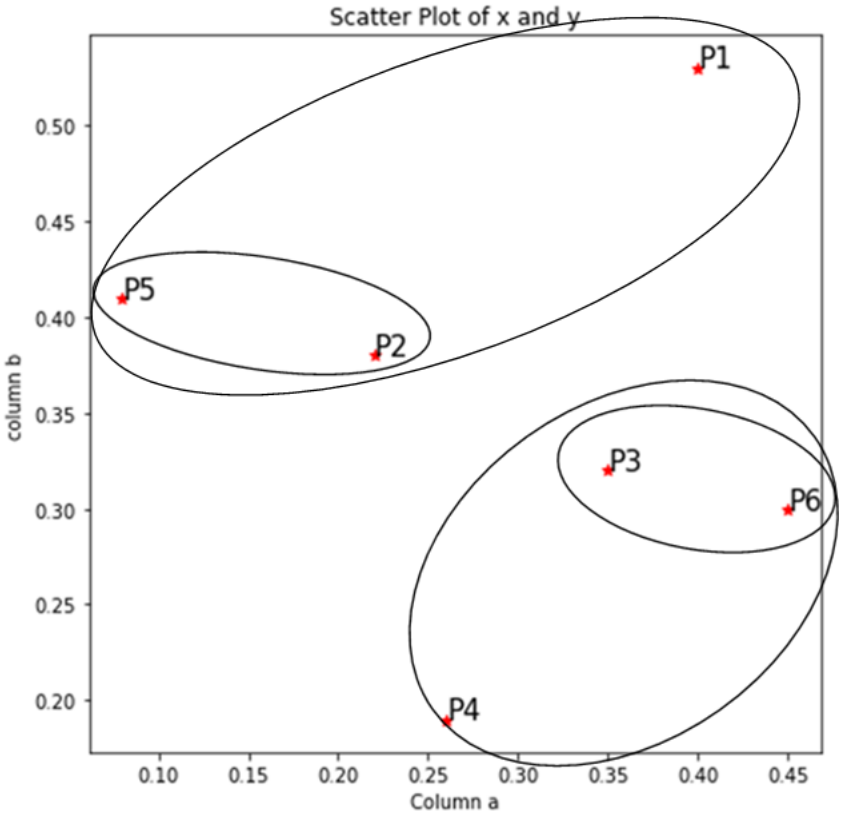


MAX(DIST(((P3,P6),P4),P1))) = MAX(DIST(((P3,P6),P1),(P4,P1)))) = MAX[(0.23,0.37)] = 0.37

MAX(DIST(((P3,P6),P4),(P4,P5)))) = MAX(DIST(((P3,P6),(P4,P5)),(P4,(P4,P5))))) = MAX[(0.39,0.29)] = 0.39

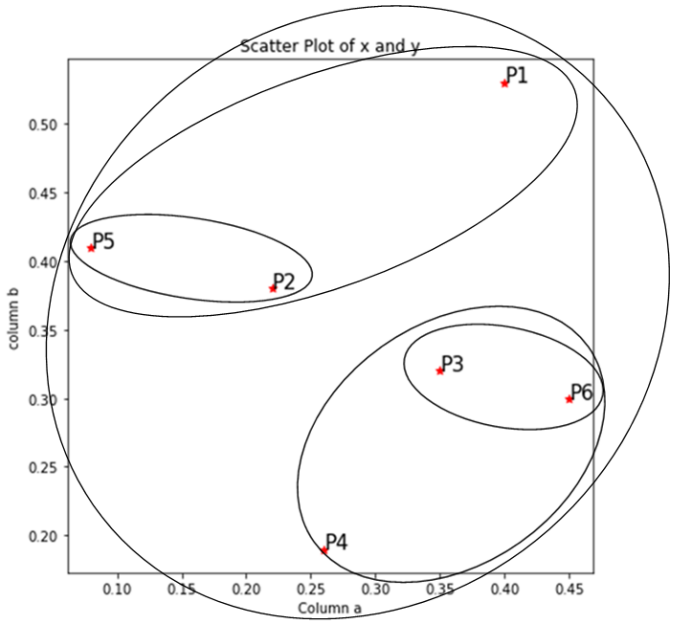
	P1	P2, P5,	P3,P6,P4
P1	0		
P2, P5	0.34	0	
P3,P6,P4	0.37	0.39	0

Grupo 4



MAX(DIST(((P2,P5),(P3,P6,P4),(P1,(P3,P6,P4))))) =
MAX(DIST(((P2,P5),(P3,P6.P4)),(P1,(P3,P5,P4))))) = MAX[(0.39,0.37)] = 0.39

	P2,P5,P1	P3,P6,P4
P2,P5,P1	0	
P3,P6,P4	0.39	0


Scatter Plot of x and y
P1
P2
P3
P4
P5
P6
column b
Column a
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

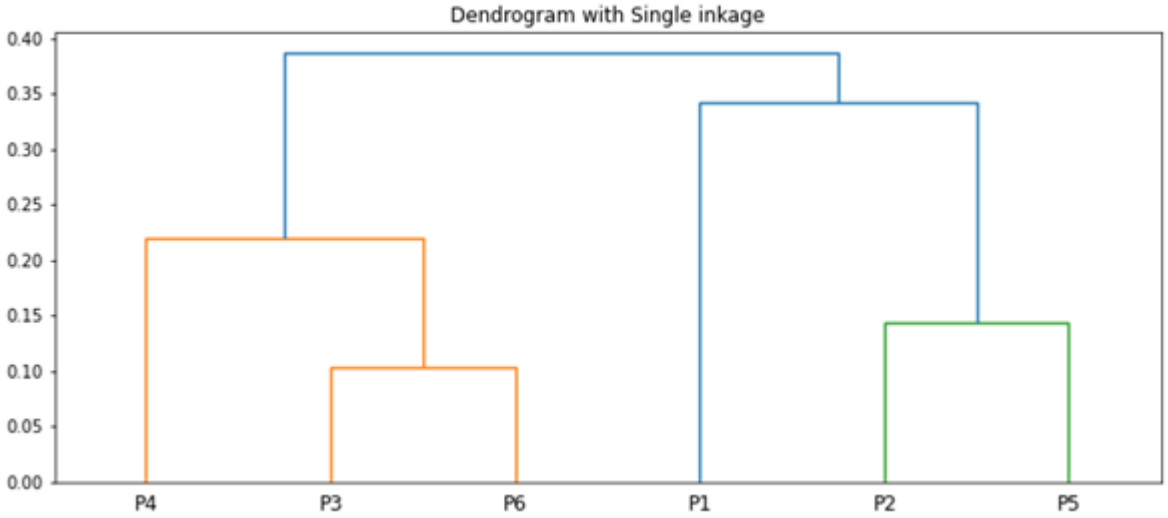
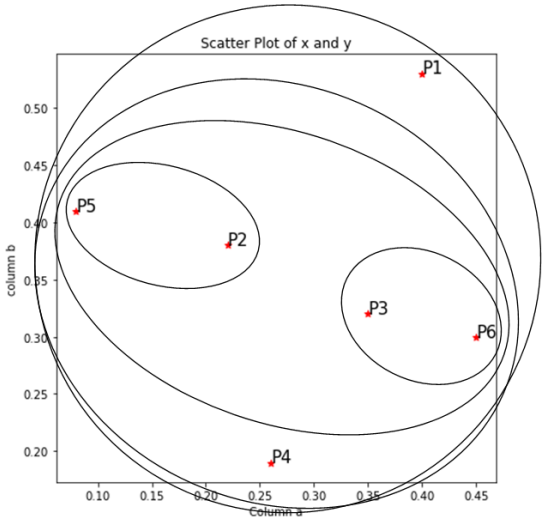
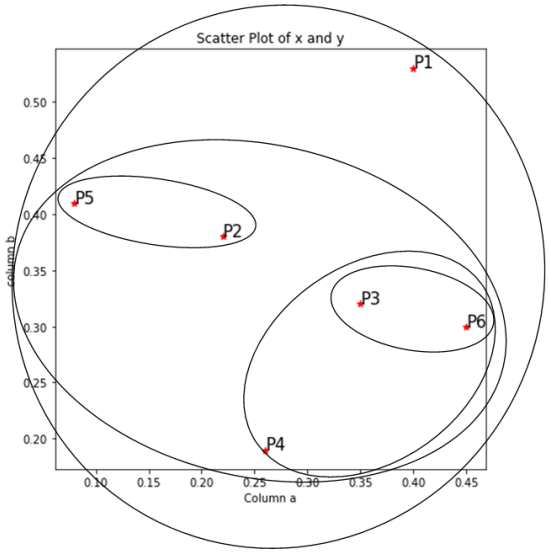

Dendrogram with Single inkage
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
P4
P3
P6
P1
P2
P5

TABLA COMPARATIVA DE CLÚSTERS OBTENIDOS:
ENLACE SIMPLE:



ENLACE PROMEDIO:



ENLACE COMPLETO:

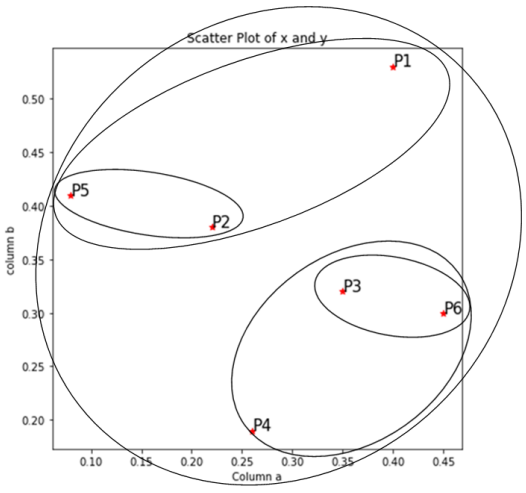
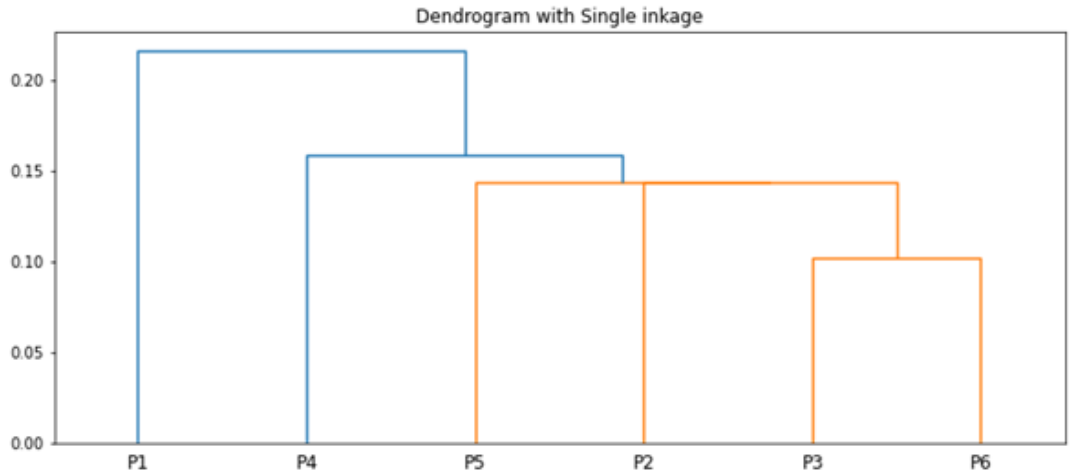
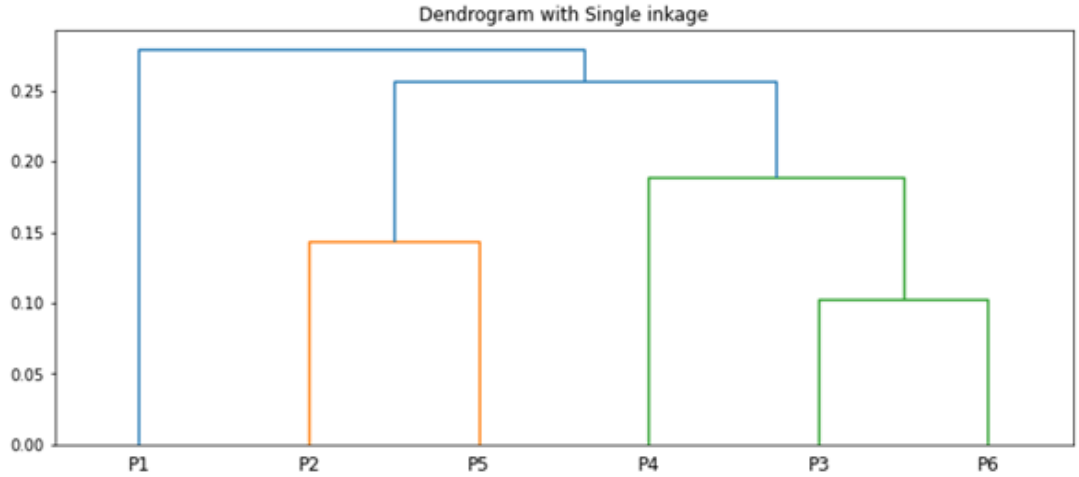


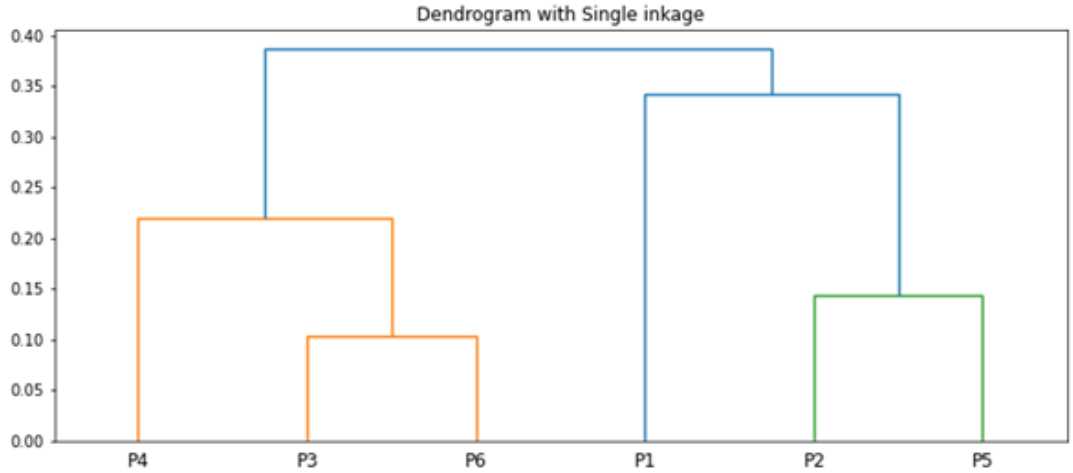
TABLA COMPARATIVA DE DENDOGRAMAS OBTENIDOS:
ENLACE SIMPLE:



ENLACE PROMEDIO:



ENLACE COMPLETO:



INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

LICENCIATURA EN CIENCIA DE DATOS

UNIDAD DE APRENDIZAJE

MINERÍA DE DATOS

CREACIÓN DE CLUSTERS - PARTE 2

NOMBRE DE LOS ALUMNOS:

DE LUNA OCAMPO YANINA

MEDINA BARRERAS DANIEL IVÁN

PROFESOR:

OCAMPO BOTELLO FABIOLA

GRUPO:

5CDM1

FECHA:

29/05/202

PARTE 2. Diseñar una matriz con datos cualitativos, cinco registros y seis características y genere los grupos.

	Tos	Fiebre	camPeso	Hemorragia	Náuseas	dolCabeza
Paciente 1	0	0	1	0	1	1
Paciente 2	1	1	0	1	0	0
Paciente 3	1	0	0	1	1	0
Paciente 4	0	0	1	1	1	0
Paciente 5	1	1	1	0	1	1

Coeficiente de Jaccard

	1	**2**	**3**	**4**	**5**
1	1				
2	0	1			
3	0.60	0.25	1		
4	0.33	0.50	0.60	1	
5	0.17	0.67	0.17	0.40	1

Pasamos a la siguiente matriz, debido al umbral, que es este caso es: 0.51

	1	**2**	**3**	**4**	**5**
1	1				
2	0	1			
3	1	0	1		
4	0	0	1	1	
5	0	1	0	0	1

Multiplicamos por sí misma, con ayuda del software obtenemos:

$$\mathbf{C} = \mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 1 & 0 & 2 & 1 & 0 \\ 0 & 2 & 0 & 0 & 1 \end{pmatrix}$$

Vemos cuántos vecinos en común hay. Obteniendo con la matriz de adyacencia lo siguiente:

	1	**2**	**3**	**4**	**5**
1	-				
2	0	-			
3	2	0	-		
4	1	0	2	-	
5	0	2	0	0	-

Podemos ver que hay grupos con el mismo número de enlaces. Escogemos uno, será el 3 con el 4.

	1	**2**	**(3, 4)**	**5**
1	-			
2	0	-		
(3, 4)	2+1	0	-	
5	0	2	0	-

Formamos un clúster ahora con el (1,3,4) por el número de enlaces contenidos. Nos quedaría:

	(1, 3, 4)	**2**	**5**
(1, 3, 4)	-		

2	0	-	
5	0	2	-

Solo nos queda el 2 y el 5 por lo que crearemos un clúster, como se muestra a continuación:

{1, 3, 4} y {2, 5}

Recordando la tabla principal, vemos que estos agrupamientos son correctos debido a que por ejemplo, el 5 tiene más cosas en común con el 2 que con los demás.
Este usa el algoritmo ROCK, que es: RObust Clustering using linKs. Sirve para conjunto de datos con atributos categóricos y booleanos. Un par de puntos se consideran vecinos si su similitud está por encima del umbral preestablecido.

- El algoritmo ROCK evalúa las distancias entre objetos utilizando el coeficiente Jaccard.
- Utiliza el parámetro θ para determinar quienes son los vecinos en cada uno de los objetos.
- Dado un punto p, un punto q es vecino de p si el coeficiente de Jaccard sim(p, q) excede el valor de θ.
- Se generan los valores de la matriz de ligas (links), la cual consiste en la evaluación de links(p, q) como el número de vecinos comunes entre los puntos p y q.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$