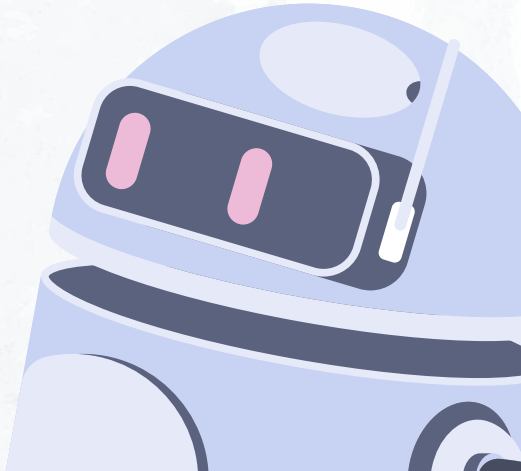
Analítica de Datos (Aprendizaje de máquina)

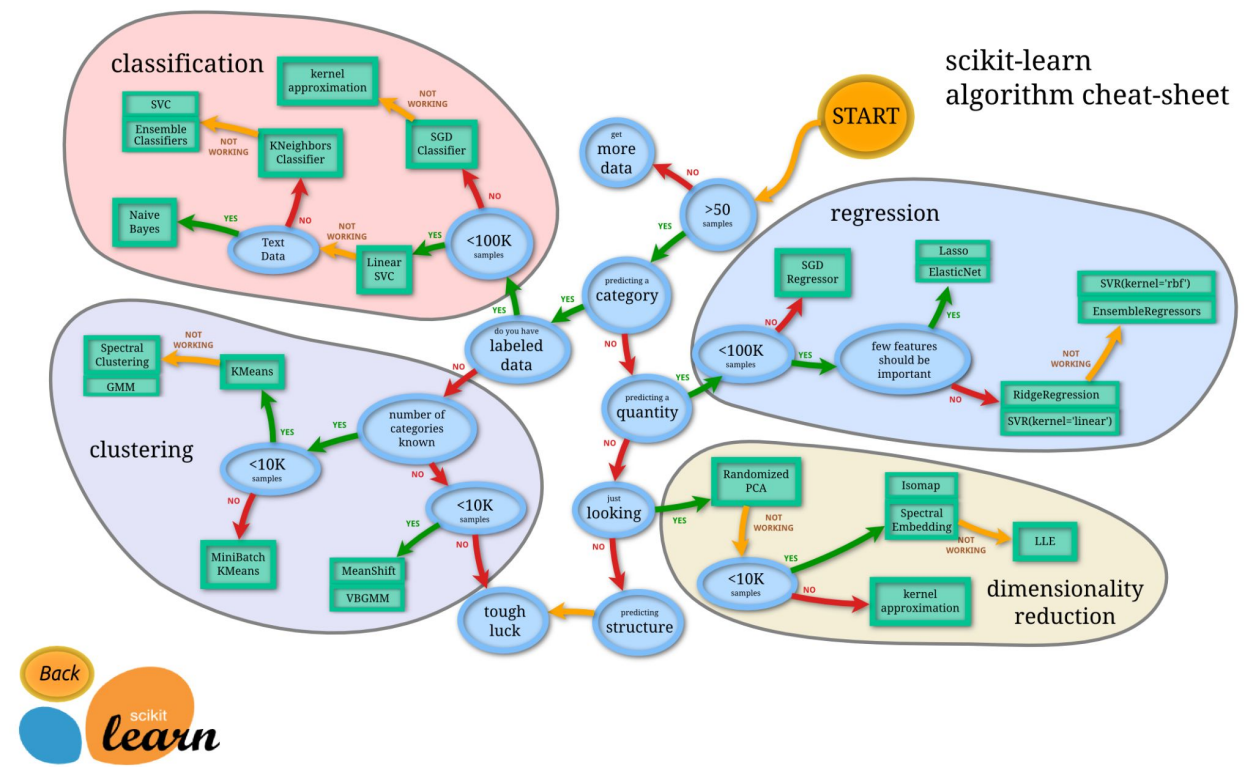


UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD NACIONAL DE COLOMBIA

Machine learning - Guideline



Análisis no supervisado

Clustering - Agrupamiento

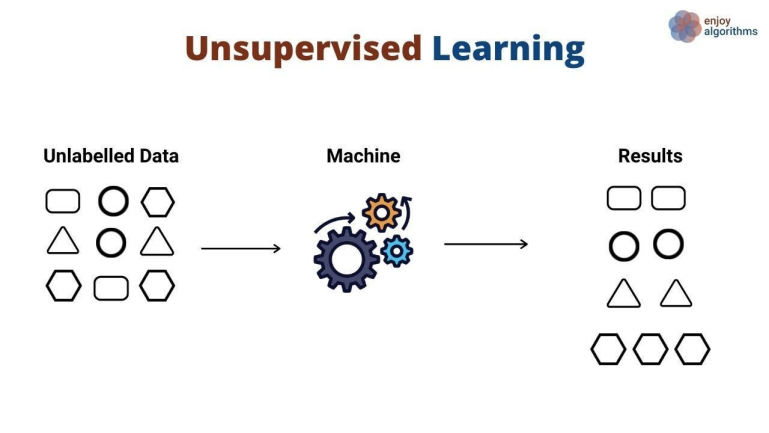
Es el proceso de organizar un conjunto de datos en grupos, de tal manera que las observaciones dentro de un grupo sean más similares entre sí que las observaciones que pertenecen a un grupo diferente.

Existen muchos métodos de agrupación y esquemas de representación de datos. No obstante, conviene señalar que ninguna técnica de agrupación es universalmente aceptada.

Y lo más importante... debemos intentar (desde el conocimiento del problema) explicar los grupos que aparecen.

Análisis no supervisado

Clustering - Agrupamiento



- En las técnicas de aprendizaje **no supervisado** no se conocen las clases a las que pertenecen las instancias (cada uno de los sujetos de la base de datos).
- De hecho, no se conoce cuántos grupos están representados en los datos o si existe alguna estructura que los agrupe.
- Se desconoce incluso si inicialmente hay algún tipo de agrupación.
- Este tipo de aprendizaje supone una exploración de los datos y el descubrimiento de relaciones entre ellos.

Ventajas ¿?

¿Cómo agrupar los datos?

Clustering - Agrupamiento

Para empezar a agrupar datos tenemos que pensar **bajo qué criterio queremos agruparlos.**

Y una forma de hacerlo sería en función de qué tanto se parecen. Para ello necesitamos **cuantificar ese criterio**, de ahí el origen del concepto de distancia y similitud entre los datos.

Estos dos conceptos son opuestos.

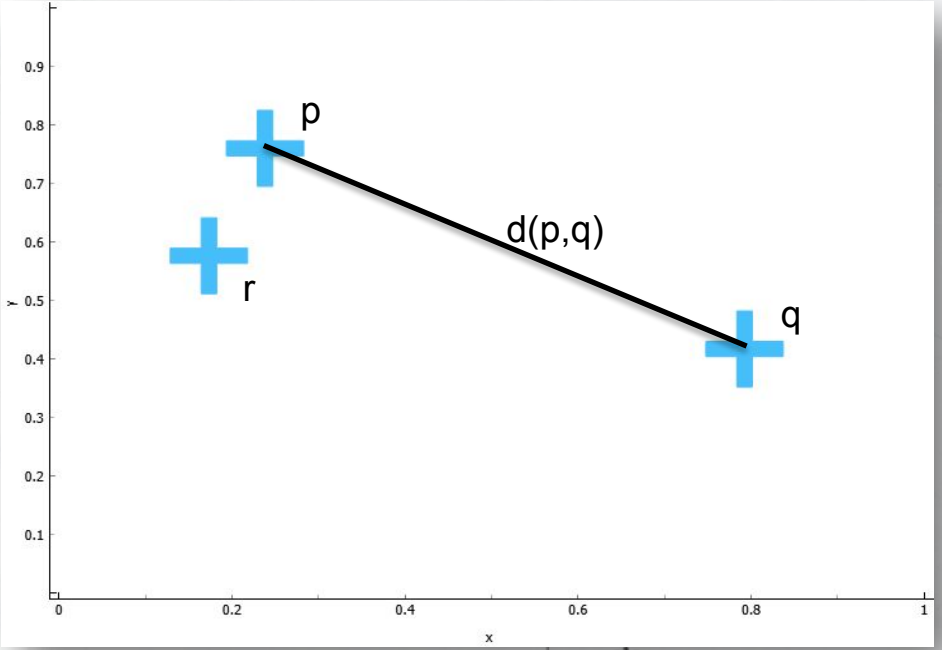
Esto quiere decir que a una menor distancia entre datos, una **mayor similitud** entre ellos.
Y a una mayor distancia, una **menor similitud**.

Las medidas de distancia son muy sensibles a:

* La distribución de los datos.
* La dimensionalidad de los datos.
* El tipo de los datos.

Distancia (característica)

Clustering - Agrupamiento



La distancia, en palabras simples, no es más que el espacio que separa dos elementos.

¿Qué debe tener una medida de distancia?

i) No-negatividad

$$d(p,q) \geq 0 \quad \forall \quad p,q$$

ii) Simetría

$$d(p,q) = d(q,p) \quad \forall \quad p,q$$

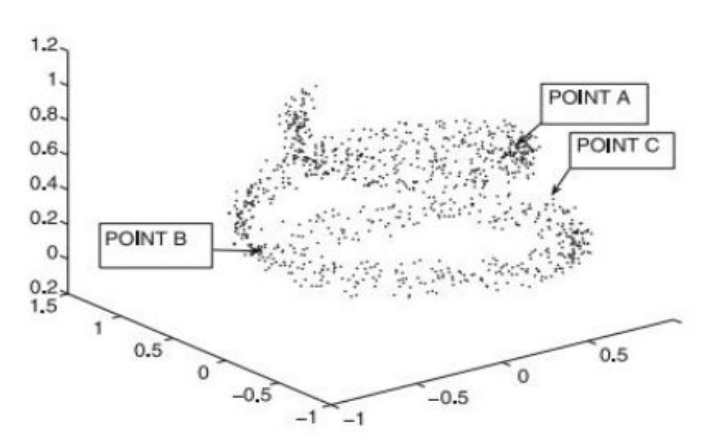
iii) Desigualdad del triángulo (desigualdad triangular)

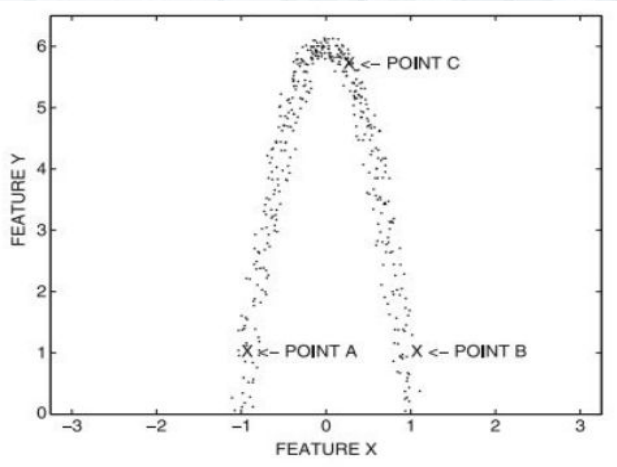
$$d(p,q) \leq d(p,r) + d(r,q) \quad \forall \quad p,q,r$$

Distancia (característica)

Clustering - Agrupamiento

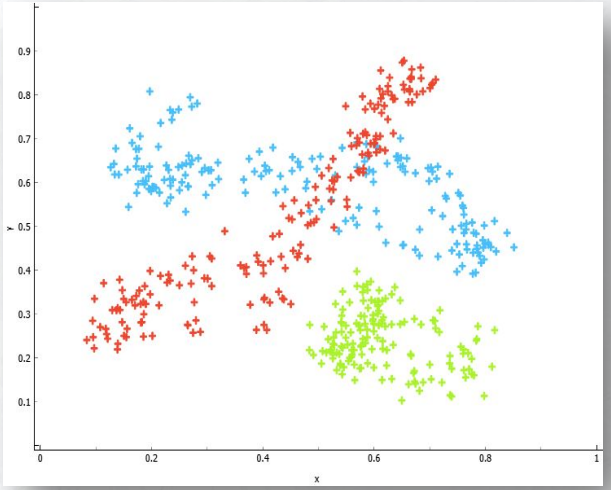
La **distancia directa** entre dos puntos no siempre es la mejor

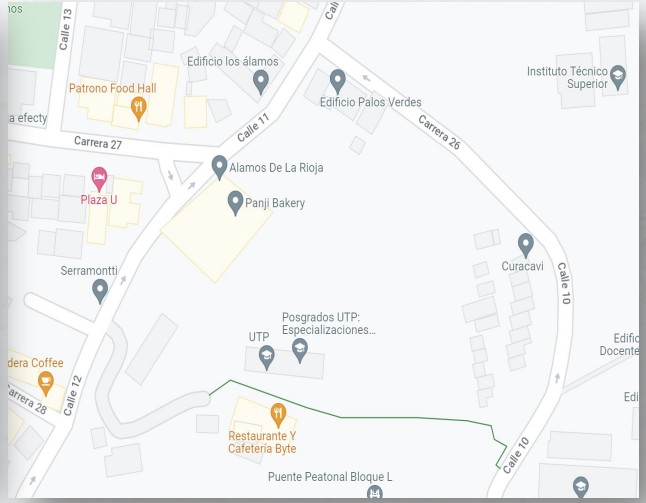




Distancia (tipos)

Clustering - Agrupamiento





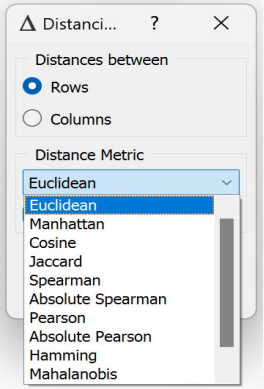
Los **datos** pueden venir representados en distintos escenarios. Además, no siempre la intención al analizarlos es la misma.

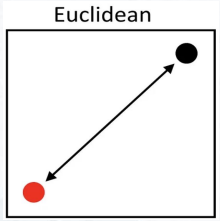
Distancia (tipos)

Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias





$$d(P,Q) = ||P-Q||_0 = \sqrt{\sum_{i=1}^{2}(p_i - q_i)^2}$$
$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

where:

$P = (p_1, p_2)$, and $Q = (q_1, q_2)$

La **distancia euclídea** representa la menor distancia entre un par de puntos.

En un sistema cartesiano representa el camino más corto para llegar desde el punto P al punto Q.

Cabe anotar que estas no son las únicas.

UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

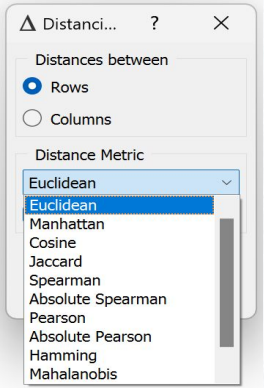
UNIVERSIDAD NACIONAL DE COLOMBIA

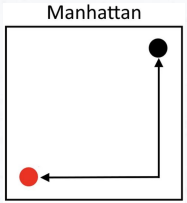
Distancia (tipos)

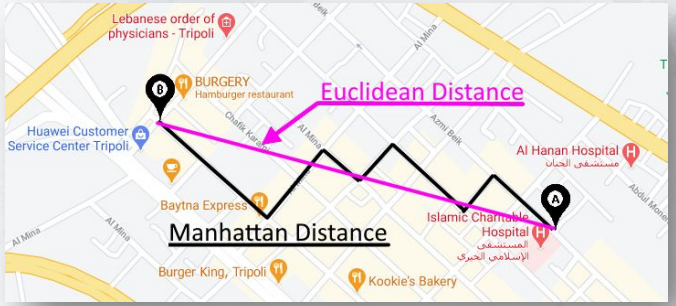
Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias







Esta métrica es muy útil para medir la distancia entre 2 calles en una ciudad.

La **distancia Manhattan** se puede medir en términos del número de manzanas que separan dos lugares diferentes.

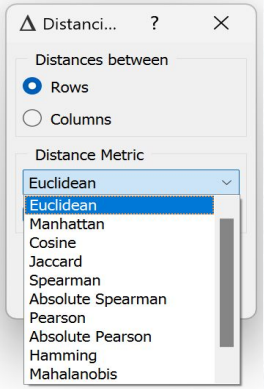
Cabe anotar que estas no son las únicas.

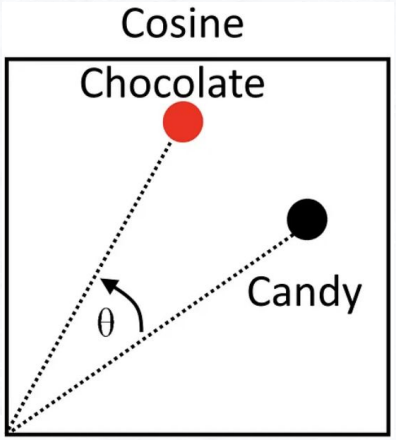
Distancia (tipos)

Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias





La **distancia coseno** es una métrica de distancia usada para medir la similitud entre dos vectores.

El cálculo de la distancia coseno se basa en el ángulo entre los vectores y comprueba la similitud entre ellos.

La distancia coseno es una técnica usada comúnmente en minería de texto, sistemas de recomendación y análisis de redes.

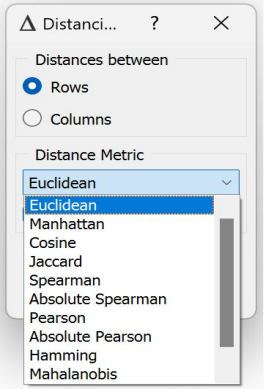
Cabe anotar que estas no son las únicas.

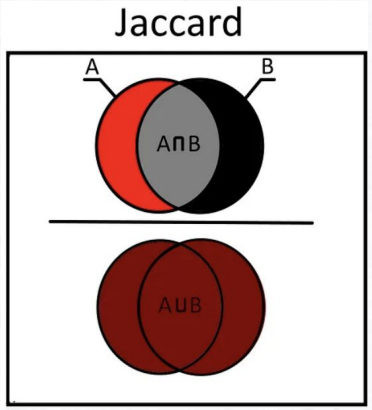
Distancia (tipos)

Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias





La **distancia de Jaccard** mide la similitud entre dos conjuntos mediante la división de la cantidad de elementos comunes entre ellos por la cantidad de elementos únicos entre cada uno.

La distancia de Jaccard es una métrica de similitud usada para medir la similitud o la diferencia entre conjuntos.

Es una métrica de similitud **no-bipolar** que devuelve un valor entre 0 y 1, donde 0 retorna una similitud inexistente y 1 retorna una similitud total.

Orange calcula la distancia de Jaccard (**1-similitud**)

$$P = \{a,b,c\} \quad Q = \{a,b,c,d,e,f,g,h\}$$

$$J = \frac{|P \cap Q|}{|P \cup Q|} = \frac{|\{a,b,c\}|}{|\{a,b,c,d,e,f,g,h\}|} = \frac{3}{8}$$

Cabe anotar que estas no son las únicas.

UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

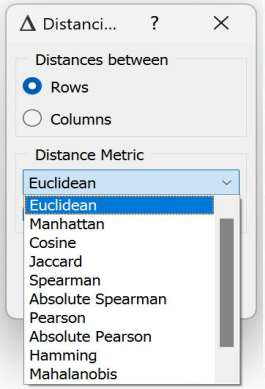
UNIVERSIDAD NACIONAL DE COLOMBIA

Distancia (tipos)

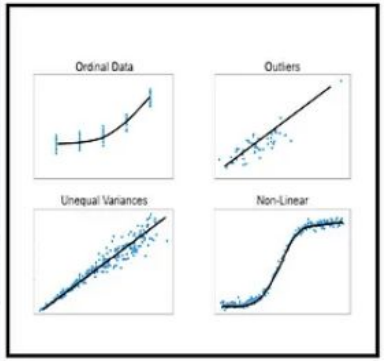
Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias



Spearman y Absolute Spearman



Cabe anotar que estas no son las únicas.

Es una medida estadística que mide la similitud entre el ranking de dos variables.

Es necesario que estas variables puedan por lo tanto ordenarse de mayor a menor.

Absolute Spearman calcula el coeficiente de spearman entre el ranking de valores absolutos de los datos

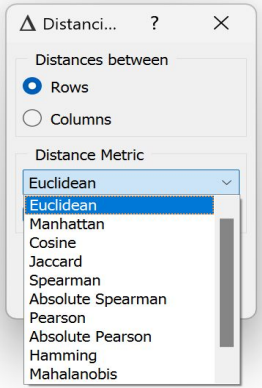
History	Rank	Geography	Rank	d
35	3	30	5	2
23	5	33	3	2
47	1	45	2	1
17	6	23	6	0
10	7	8	8	1
43	2	49	1	1
9	8	12	7	1
6	9	4	9	0
28	4	31	4	0

Distancia (tipos)

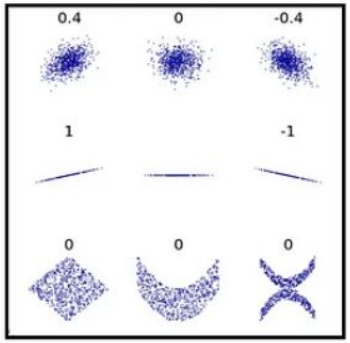
Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias



Pearson y Absolute Pearson

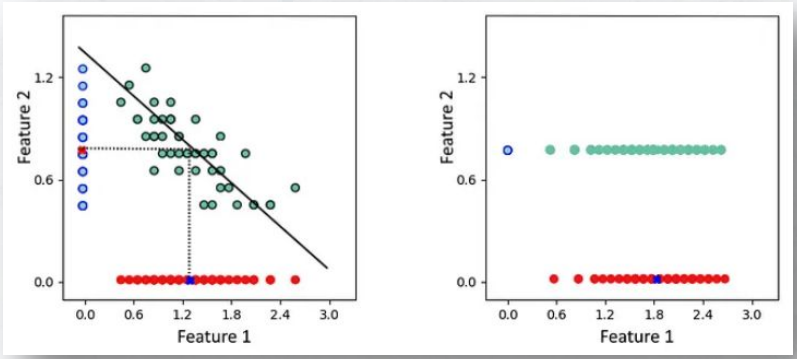


Cabe anotar que estas no son las únicas.

La **correlación de Pearson** mide la relación lineal entre dos variables.

Está entre -1 y 1, indicando positiva o negativamente la asociación entre las variables.

Absolute Pearson calcula el coeficiente de Pearson entre el ranking de valores absolutos de los datos.



UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

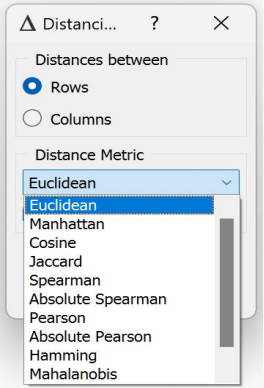
UNIVERSIDAD NACIONAL DE COLOMBIA

Distancia (tipos)

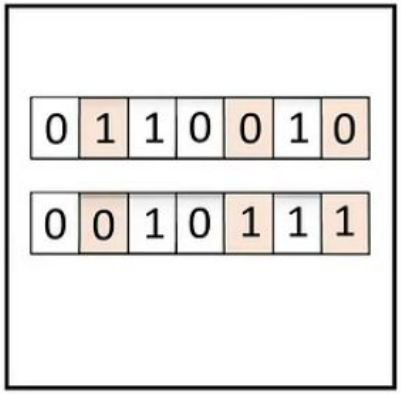
Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias



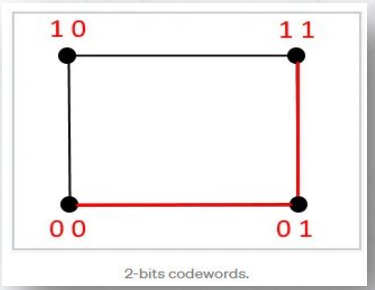
Hamming

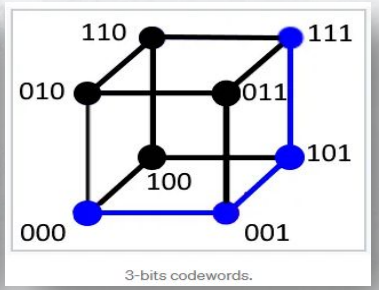


La **distancia de hamming** es igual al número de dígitos donde difieren dos palabras clave de la misma longitud.

En el mundo binario, es igual al número de bits diferentes entre dos mensajes binarios.

H(100001, 010001) = 2
H(110, 111) = 1





Cabe anotar que estas no son las únicas.

UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

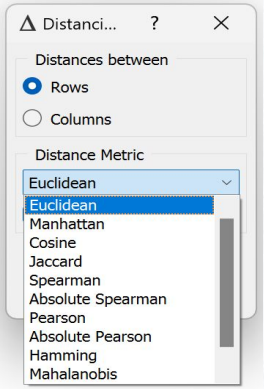
UNIVERSIDAD NACIONAL DE COLOMBIA

Distancia (tipos)

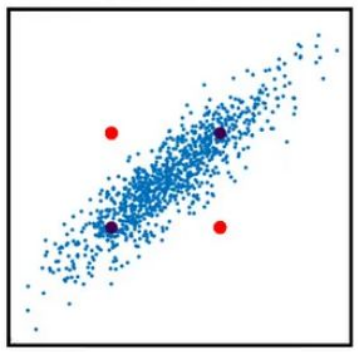
Clustering - Agrupamiento

Veamos los tipos de distancias que permite calcular Orange.

Distancias



Mahalanobis

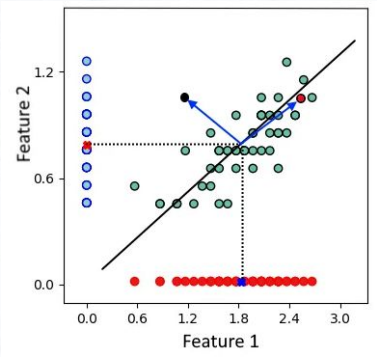


Cabe anotar que estas no son las únicas.

La **distancia de Mahalanobis** se usa para calcular la distancia entre dos vectores en un espacio multivariante.

Esta distancia **involucra una matriz de covarianza**, la cual refleja la correlación lineal entre los diferentes atributos.

Útil si la distancia euclídea no representa la realidad buscada



UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD NACIONAL DE COLOMBIA

Distancia (tipos)

Clustering - Agrupamiento

Existen muchas **más distancias** para variables categóricas

Medidas para datos dicotómicos

$X_i \setminus X_j$	1	0	Totales
1	a	b	a + b
0	c	d	c + d
Totales	a + c	b + d	m = a + b + c + d

Medida de Ochiai → $\frac{a}{\sqrt{(a+b)(a+c)}}$

Medida Φ → $\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$

Medida de Russell y Rao → $\frac{a}{a+b+c+d} = \frac{a}{m}$

Medida de Parejas simples → $\frac{a+d}{a+b+c+d} = \frac{a+d}{m}$

Medida de Jaccard → $\frac{a}{a+b+c}$

Medida de Dice → $\frac{2a}{2a+b+c}$

Medida de Rogers-Tanimoto → $\frac{a+d}{a+d+2(b+c)}$

Distancia (operaciones)

Clustering - Agrupamiento

Algunas **distancias** pueden dar como resultado valores muy dispares. Se puede aplicar una normalización a las distancias para hacerlas proporcionales.

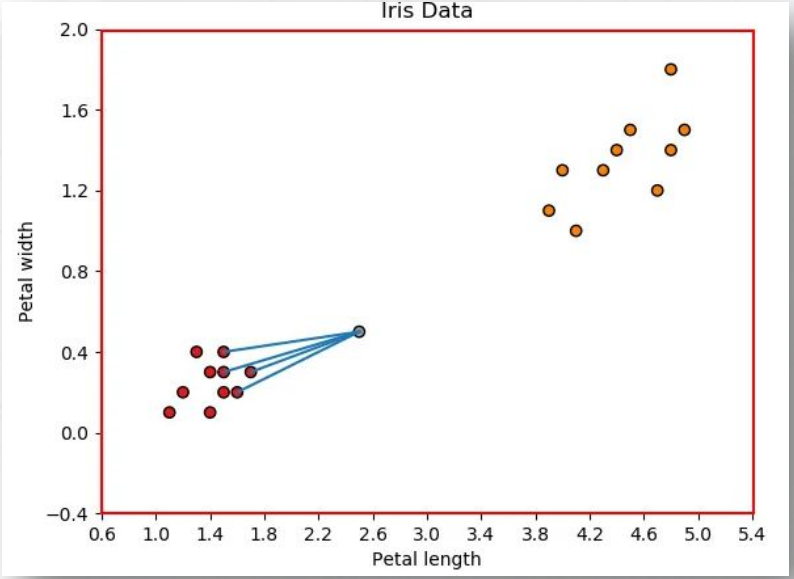
	buey1	buey2	burro1	burro2	cisne1	cisne2
buey1		35,678	43,471	38,634	43,177	45,251
buey2	35,678		48,293	45,393	48,123	49,334
burro1	43,471	48,293		40,886	49,104	50,097
burro2	38,634	45,393	40,886		45,814	46,494
cisne1	43,177	48,123	49,104	45,814		27,880
cisne2	45,251	49,334	50,097	46,494	27,880	

	buey1	buey2	burro1	burro2	cisne1	cisne2
buey1		0,645	0,786	0,698	0,781	0,818
buey2	0,645		0,873	0,821	0,870	0,892
burro1	0,786	0,873		0,739	0,888	0,906
burro2	0,698	0,821	0,739		0,828	0,841
cisne1	0,781	0,870	0,888	0,828		0,504
cisne2	0,818	0,892	0,906	0,841	0,504	

Vecinos

Clustering - Agrupamiento

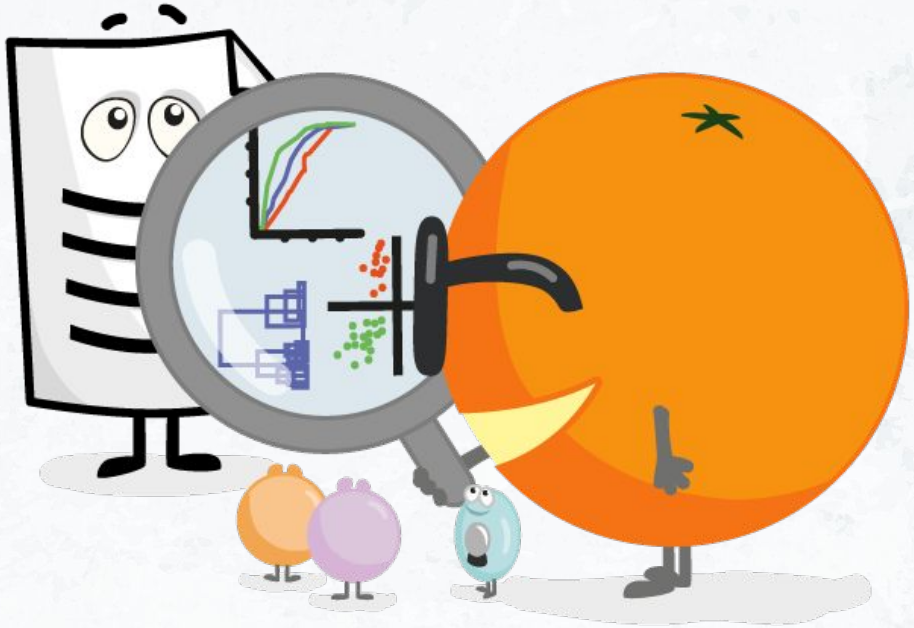
Un **vecino** podemos pensar que es un dato que se parece al dato original, se encuentra cercano o tiene una o varias características similares a las de dicho dato.



Gracias a las medidas de distancia podemos definir **qué datos serán vecinos de cuáles**.

No todas las medidas son igual de eficaces para escoger vecinos

Ejercicios prácticos

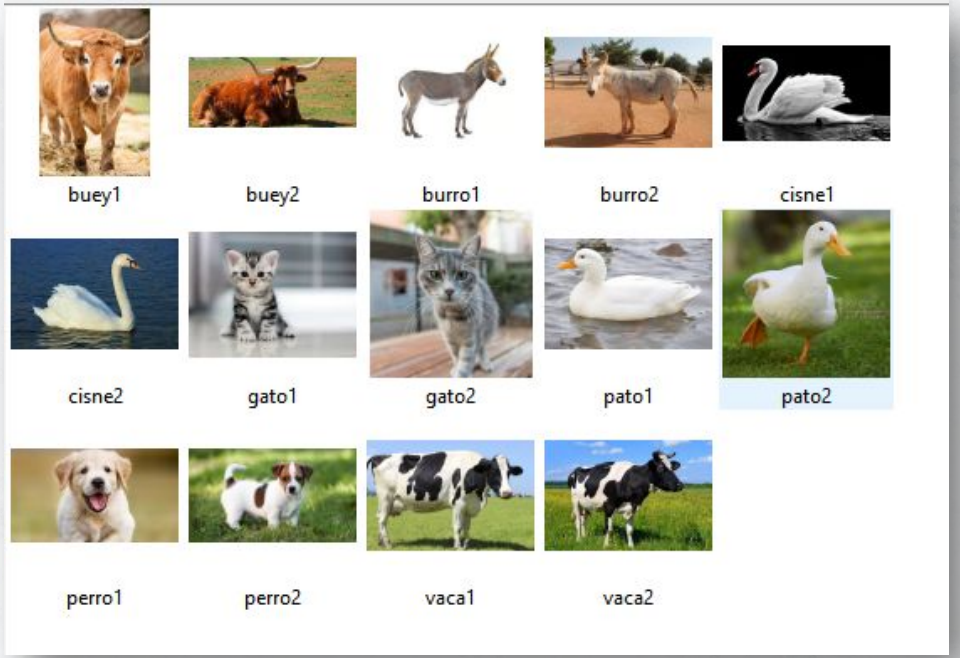


UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Caso práctico

Clustering - Agrupamiento

Distancia entre imágenes

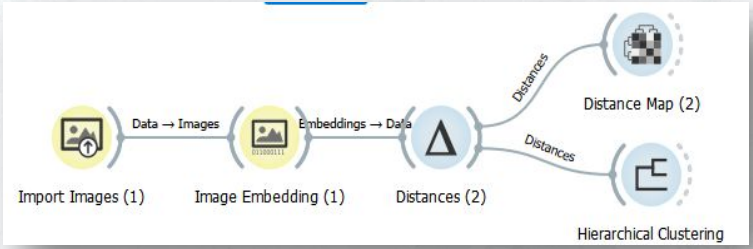


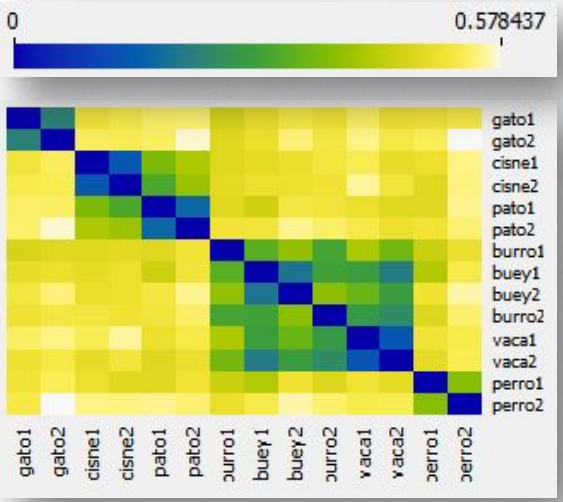
UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

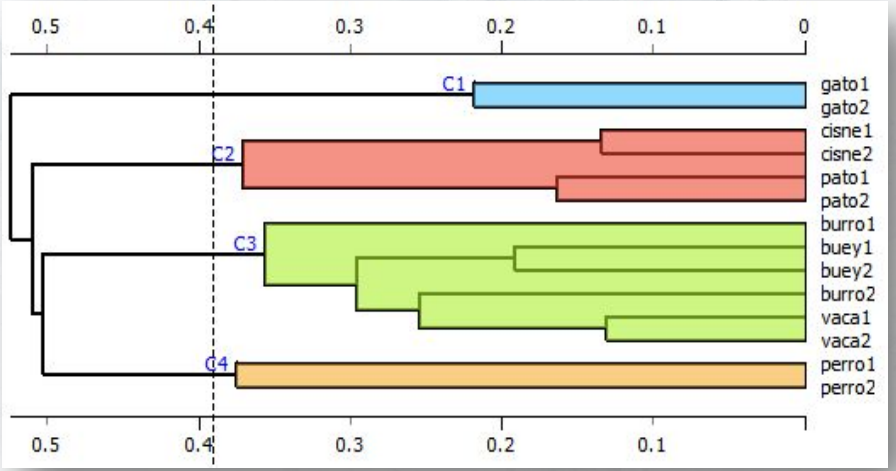
Caso práctico

Clustering - Agrupamiento

Distancia coseno y agrupamiento



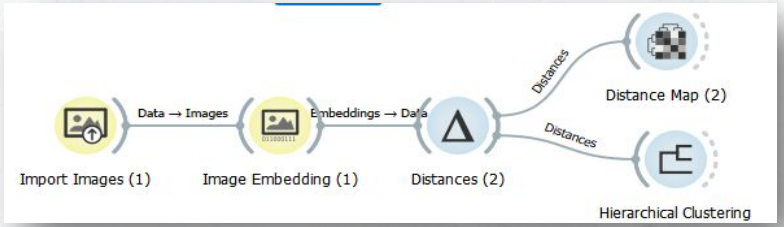


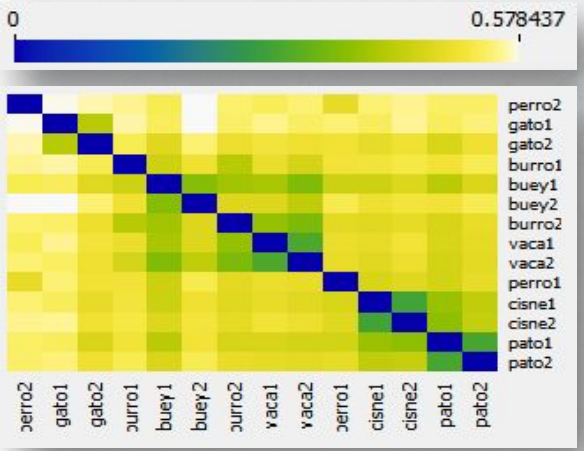


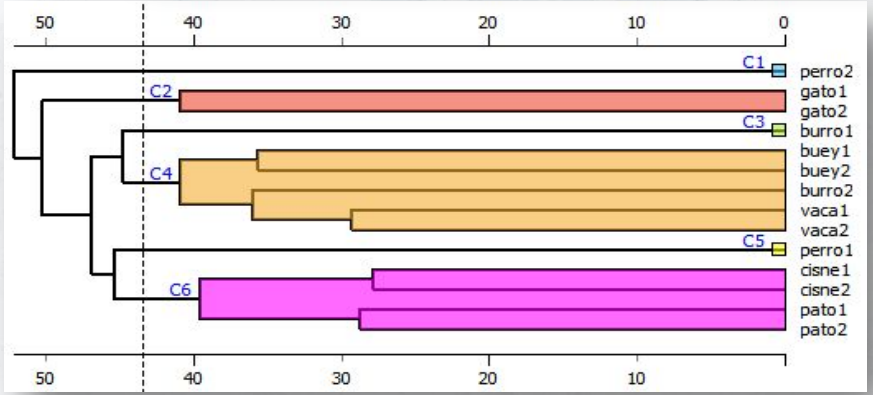
Caso práctico

Clustering - Agrupamiento

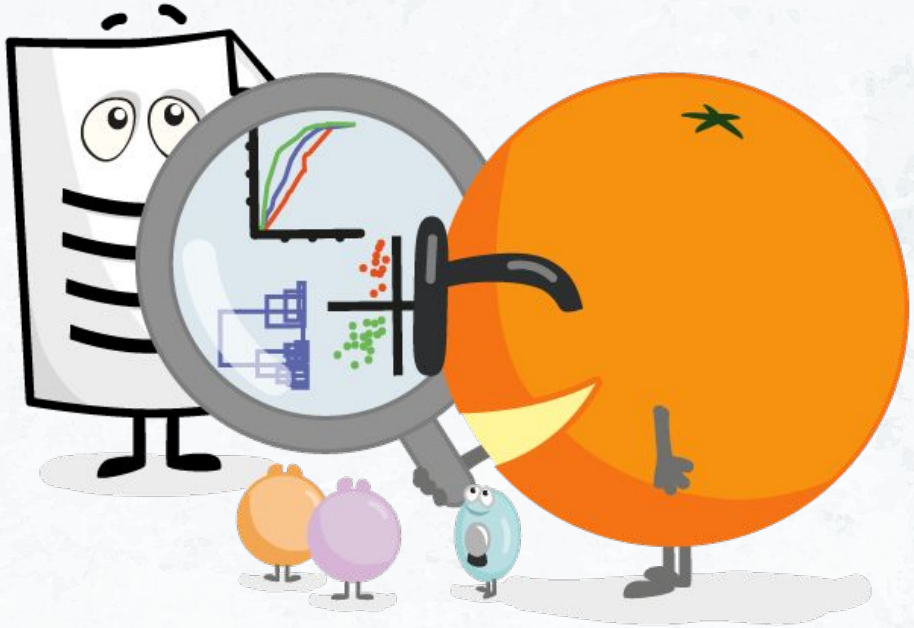
Distancia euclídea y agrupamiento







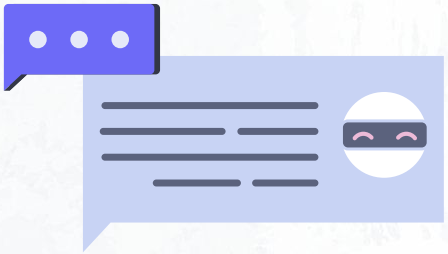
Ejercicios prácticos

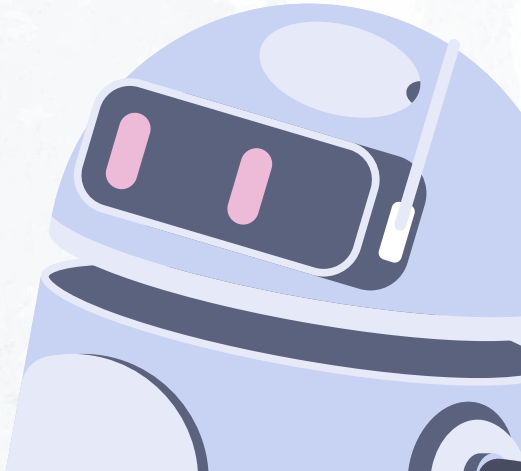


UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Gracias !

dfcollazosh@unal.edu.co





UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD
NACIONAL
DE COLOMBIA