

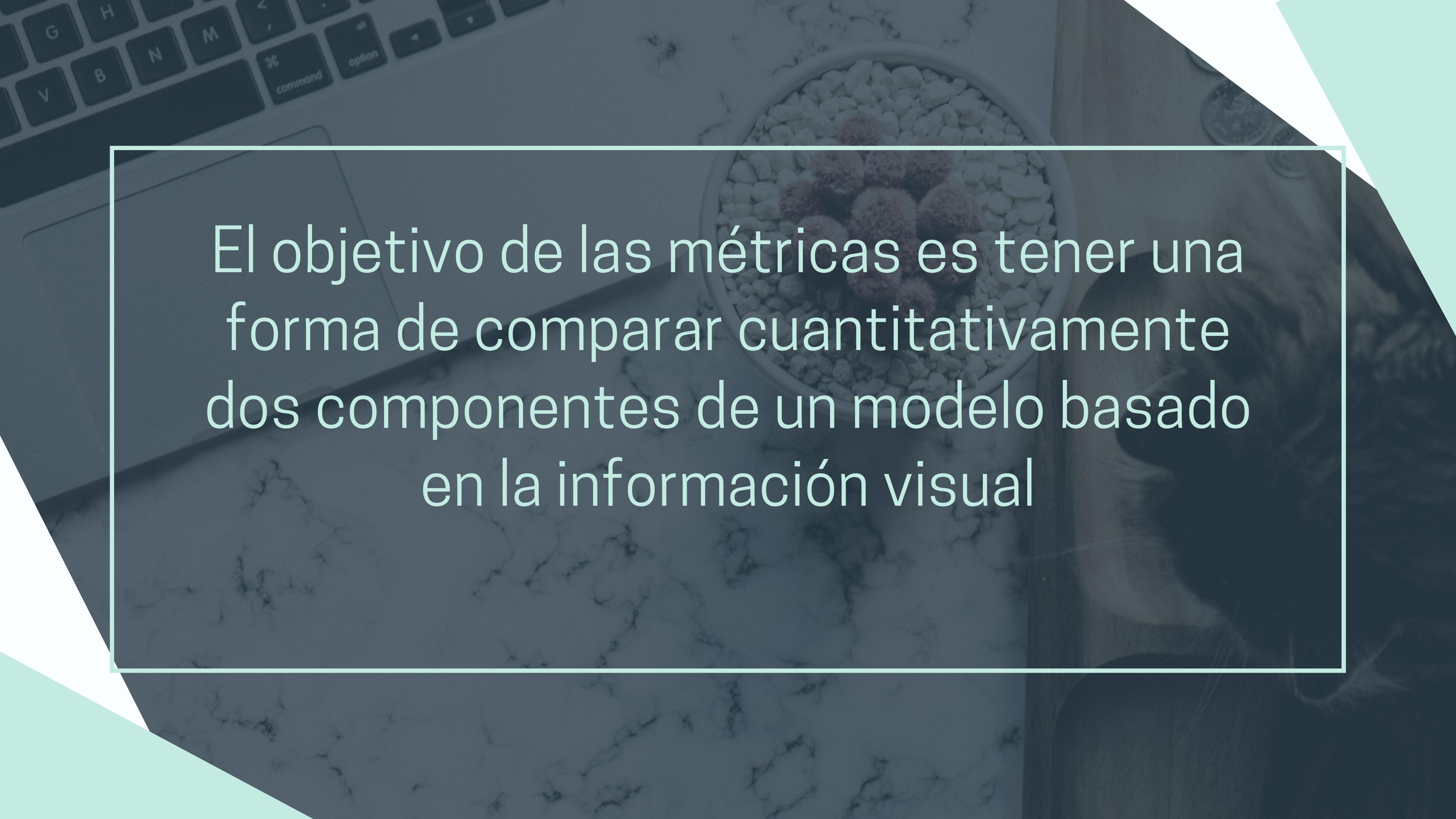
Métricas de Evaluación

CLUSTERING

Un algoritmo de clustering tiene como objetivo agrupar los objetos de un dataset según su similaridad , de forma que los datos que hay dentro de un grupo (cluster) sean más parecidos que aquellos que estan en grupos distintos.

METRICAS DENTRO DE CLUSTERING

Permiten determinar la similitud entre dos instancias, si se ve cada componente de un modelo de minería de datos, como un grupo de instancias asociadas a reglas generadas por la técnica árbol de decisión como, por ejemplo, homogeneidad o dispersión asociada a cada componente.



El objetivo de las métricas es tener una forma de comparar cuantitativamente dos componentes de un modelo basado en la información visual



MÉTRICA DE DISTANCIA PARA MAPAS SOM

Esta métrica representa una adaptación de la medida de distancia entre vectores de datos.

MEDIDA DE DISTANCIA

Debe ser simétrica y su valor mínimo es cero dada la condición que para dos instancias comparables resultan iguales si $x = y$



MÉTRICA DE MINKOWSKI

$$d(x_i, x_j) = \sqrt{g} \left(|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \dots + |x_{ip} - x_{jp}|^g \right)^{\frac{1}{g}}$$

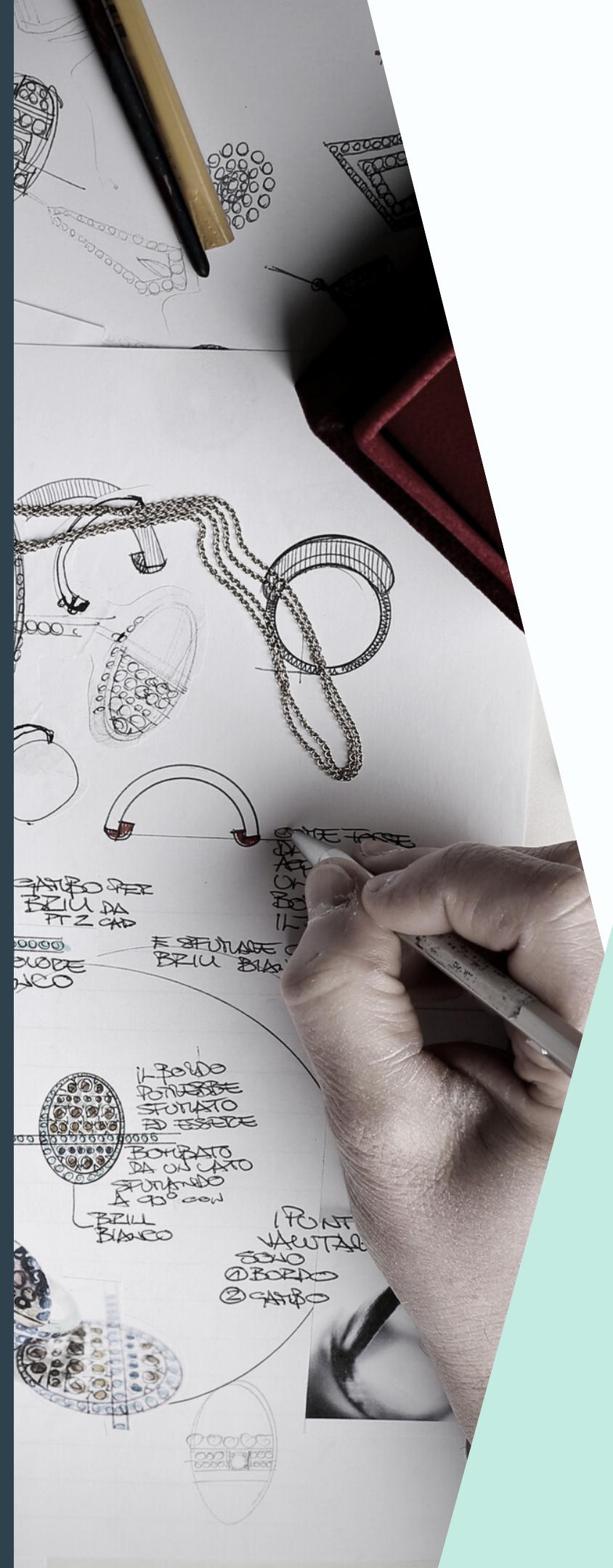
Donde:

p dimensiones

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$$

$$x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$$

Euclides con $g = 2$, Manhattan con $g = 1$, y Chebychev con $g = \infty$





LA DISTANCIA PONDERADA POR EL PESO DE CADA ATRIBUTO QUEDA COMO

$$d(x_i, x_j) = \sqrt{w_1 |x_{i1} - x_{j1}|^g + w_2 |x_{i2} - x_{j2}|^g + \dots + w_p |x_{ip} - x_{jp}|^g}$$

Donde:
 $w_i \in [0, \infty]$

PARA CONJUNTOS

atributos con valores numéricos el promedio aritmético de los valores de las instancias para cada neurona de ambos mapas (x_v e Y_v)



atributos con valores no-numéricos se utiliza el valor que más se repite es decir la moda

los pesos (w) de los atributos se utiliza un arreglo w aleatorio que posteriormente es mejorado mediante el cálculo de Best Matching Units

MÉTRICA DE SIMILITUD PARA MAPAS SOM

Sirve para comparar dos componentes representados como vectores x_i y_j , en función de su nivel de similitud denotada por $S(X_i, Y_j)$.

Esta función debe ser simétrica, $S(X_i, Y_j) = S(X_j, Y_i)$

- Si X_v e Y_v son dos vectores de instancias p dimensionales, entonces la función $S(X_v, Y_v)$ mide la similitud entre estos dos vectores
- Función de similitud dicotómica, $-1 \leq S(X_v, Y_v) \leq 1$

Medida de correlación de Pearson

$$S(X_v, Y_v) = \frac{(X_v - \bar{x}) \cdot (Y_v - \bar{y})}{\|X_v - \bar{x}\| \cdot \|Y_v - \bar{y}\|}$$

Donde:

x e y son los valores promedio de los vectores Xv e Yv, respectivamente

PARA CONJUNTOS

Dados dos vectores X_v e Y_v de $(n \cdot m)$ dimensiones, en que cada neurona $X_{vi} \in X_v$ con t instancias y p atributos, y cada neurona $Y_{vi} \in Y_v$ con k instancias y p atributos, la expresión general de la similitud a través del cálculo de correlación entre dos neuronas X_{vi} e Y_{vi}

$$S(X_{vi}, Y_{vi}) = \frac{(X_{vi} - \bar{X}) * (Y_{vi} - \bar{Y})}{\|X_{vi} - \bar{X}\| * \|Y_{vi} - \bar{Y}\|}$$

Ejemplo

Usemos lo anterior para calcular la matriz de correlación tomando como vector x (-2,-1,0,1,2) y como vector y (4,1,3,2,0)

```
import numpy as np
import matplotlib.pyplot as pltx_
x_simple = np.array([-2, -1, 0, 1, 2])
y_simple = np.array([4, 1, 3, 2, 0])
my_rho = np.corrcoef(x_simple, y_simple)

print(my_rho)

[[ 1. -0.7]
 [-0.7  1. ]]
```

Para este ejemplo obtuvimos una diagonal con valor de 1, lo que nos indica que la correlacion es positiva, las variables estan distribuidas de forma aproximada, su asociacion es lineal y no hay datos atipicos

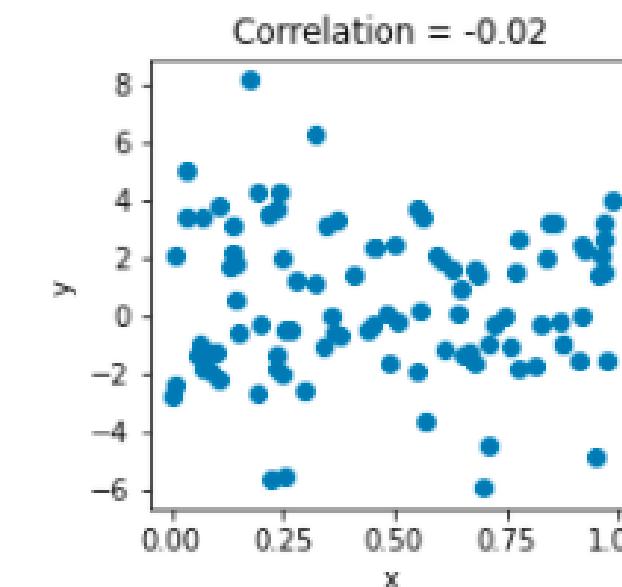
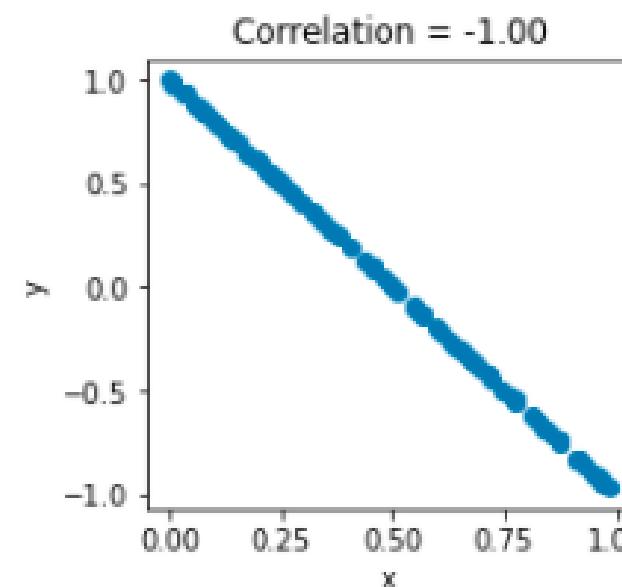
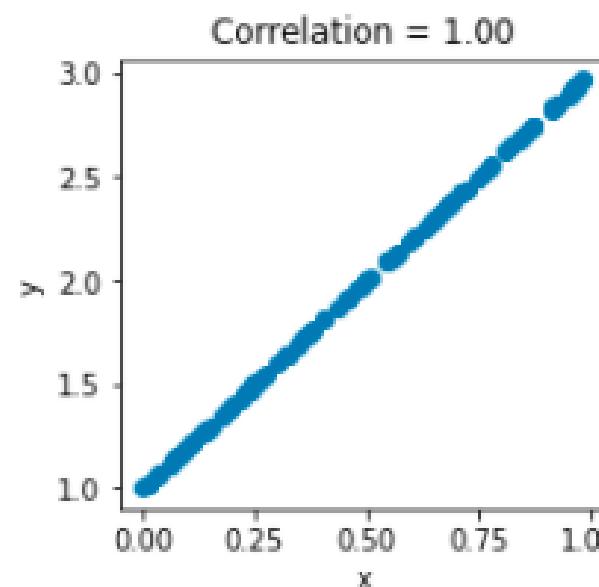
A continuación mostraremos como se visualizaría la correlación en diferentes casos

```
seed = 13
rand = np.random.RandomState(seed)

x = rand.uniform(0,1,100)
x = np.vstack((x,x*2+1))
x = np.vstack((x,-x[0,:]*2+1))
x = np.vstack((x,rand.normal(1,3,100)))

rho = np.corrcoef(x)

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(12, 3))
for i in [0,1,2]:
    ax[i].scatter(x[0,:],x[1+i,:])
    ax[i].title.set_text('Correlation = ' + "{:.2f}".format(rho[0,i+1]))
    ax[i].set(xlabel="x",ylabel="y")
fig.subplots_adjust(wspace=.4)
plt.show()
```



PREGUNTAS

- ¿Cuál es el objetivo de un algoritmo de clustering?
R-Agrupar los objetos de un dataset según su similaridad.
- ¿Cuál es el objetivo de las métricas?
R-.Es tener una forma de comparar cuantitativamente dos componentes de un modelo basado en la información visual.
- ¿Que representa la métrica de distancia para mapa som?
R-.Una adaptación de la medida de distancia entre vectores de datos
- ¿Cómo debe ser la medida de distancia?
R-.Debe ser simétrica y su valor mínimo es cero dada la condición que para dos instancias comparables resultan iguales si $x = y$.
- ¿Pará que sirve una métrica de similitud par mapas som?
R-.Para comparar dos componentes representados como vectores x_i , y_j , en función de su nivel de similitud denotada por $S(X_i, Y_j)$.

REFERENCIAS

- Correlacion lineal con python. (s/f). Cienciadedatos.net. Recuperado el 6 de septiembre de 2021, de <https://www.cienciadedatos.net/documentos/pystats05-correlacion-lineal-python.html>
- (S/f-a). Conicyt.cl. Recuperado el 6 de septiembre de 2021, de <https://scielo.conicyt.cl/pdf/ingeniare/v28n4/0718-3305-ingeniare-28-04-596.pdf>
- (S/f-b). Cs.us.es. Recuperado el 6 de septiembre de 2021, de <http://www.cs.us.es/~fsancho/?e=230>
- (S/f-c). Questionpro.com. Recuperado el 6 de septiembre de 2021, de <https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/>