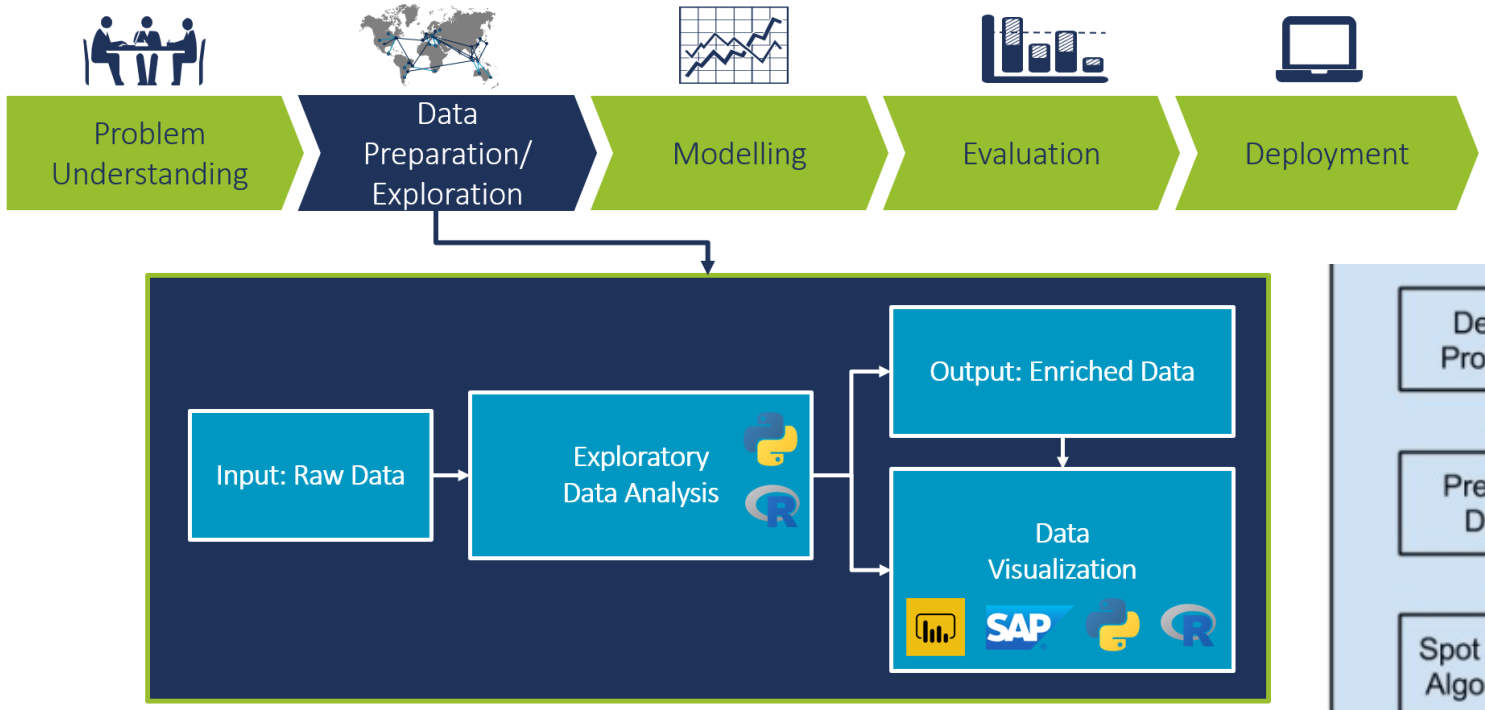
Universidad Nacional Autónoma de México
Facultad de Ingeniería

Selección de características

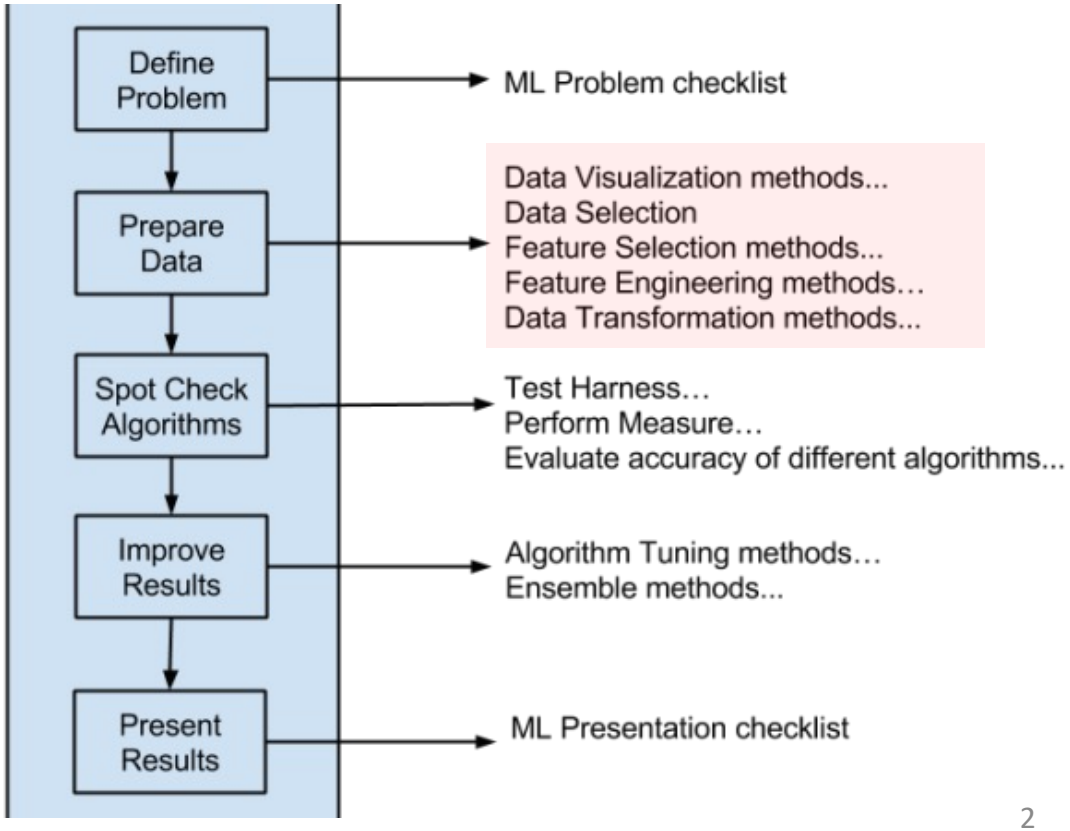
Guillermo Molero-Castillo
guillermo.molero@ingenieria.unam.edu

Septiembre, 2022

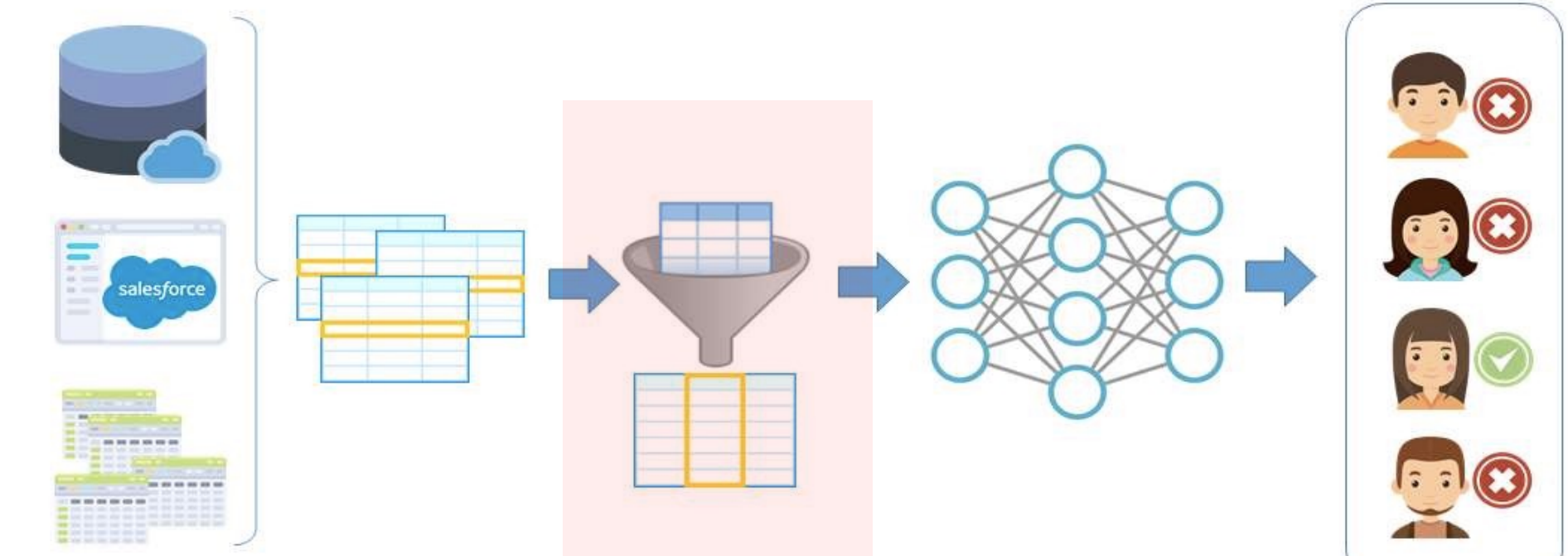
Selección de características



Selección de características (variables)



Selección de características



Fuentes de datos

Selección de variables

Aprendizaje automático

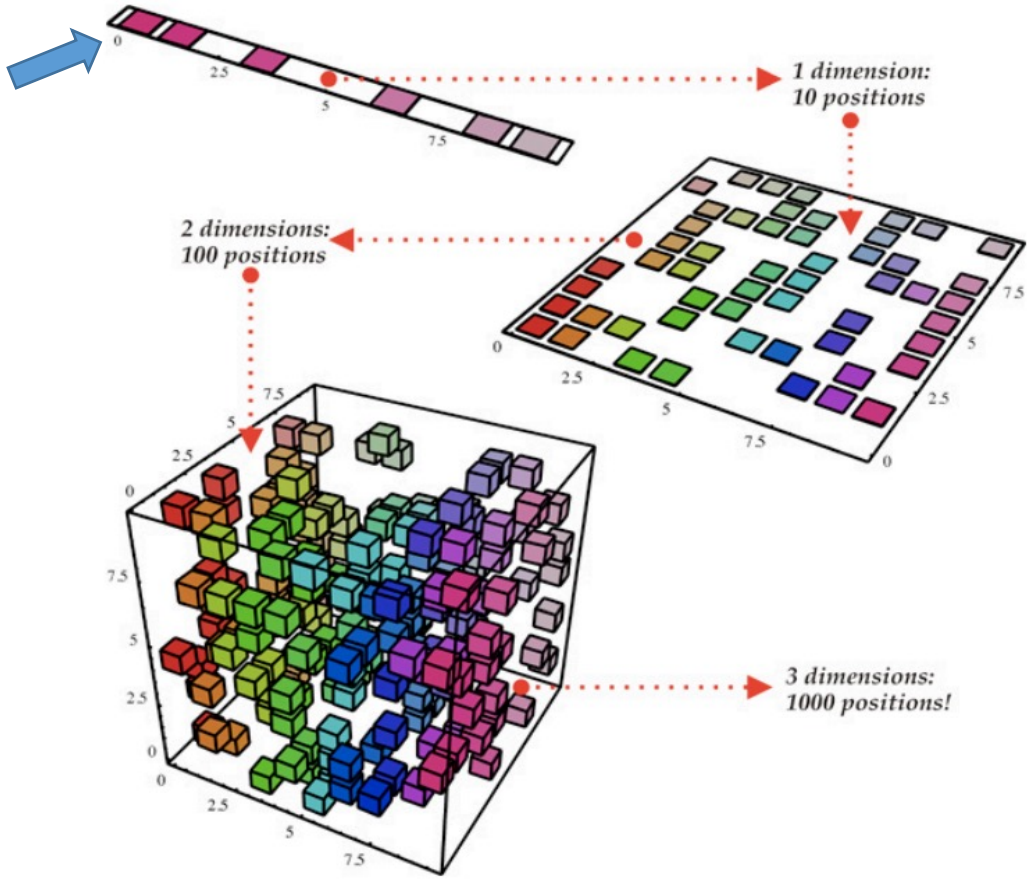
Resultados

Selección de características

A menudo hay demasiadas **variables,** en función de las cuales se condiciona el resultado final de un modelo.

- Cuanto mayor es el número de variables, es más complejo visualizar los datos y más complejo aún trabajar con éstos.
- No es ilógico pensar que entre más variables (características/atributos) se tenga en cuenta será mejor.
- Sin embargo, esto es un **error común** que no se debe cometer.

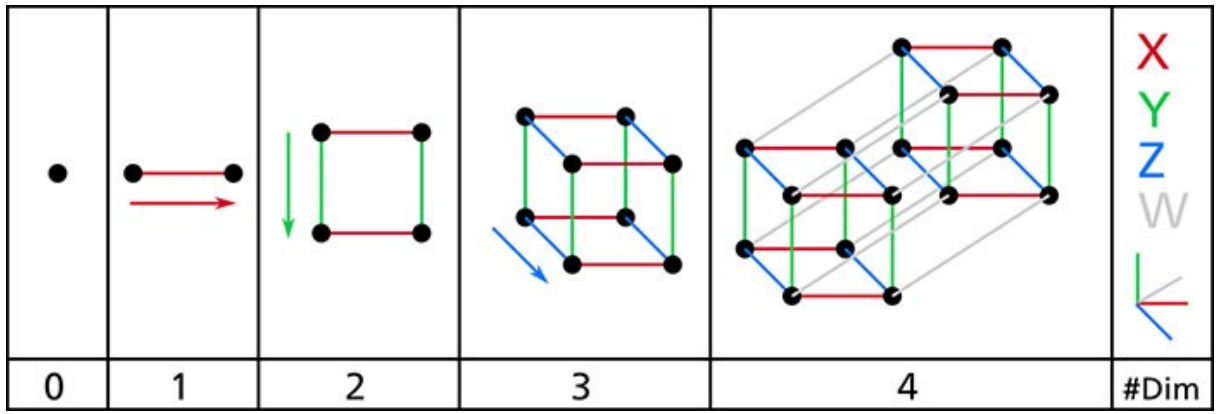
	F_1
E_1	3
E_2	2
E_3	5
E_4	10
E_5	3
E_6	4
E_7	6
E_8	7
E_9	9
E_{10}	1

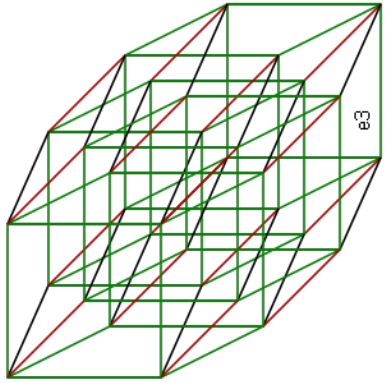


Dimensionalidad de datos

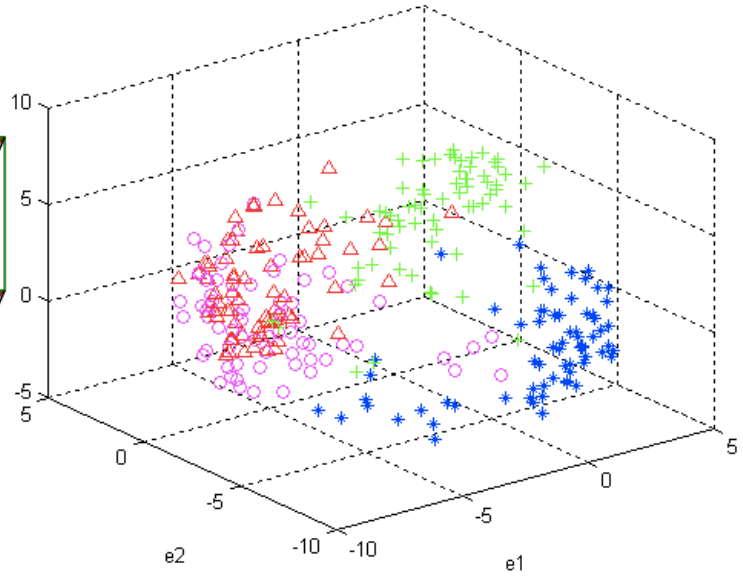
Maldición de la dimensionalidad

- La **maldición de la dimensionalidad de datos** es un problema que se puede presentar si se quiere tener en cuenta todas las características (variables) posibles en un sistema.
- Esta **maldición** hace referencia al aumento exponencial de la dimensionalidad de datos.



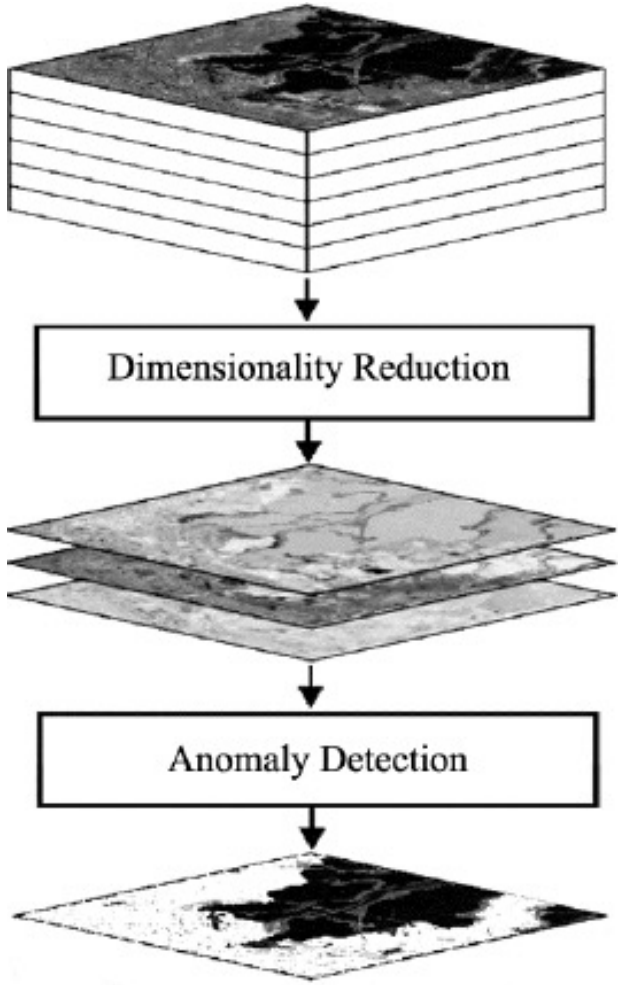


5 Dimensiones



Maldición de la dimensionalidad

- En general, si la mayoría de las **variables están correlacionadas**, entonces algunas de éstas son redundantes.
- Aquí es donde se utilizan estrategias para la reducción de la dimensionalidad de datos.
- Esta **reducción de la dimensionalidad** es el proceso de aminorar el número de variables mediante la obtención de alguna función de puntuación, que generalmente mide la relevancia de las características.

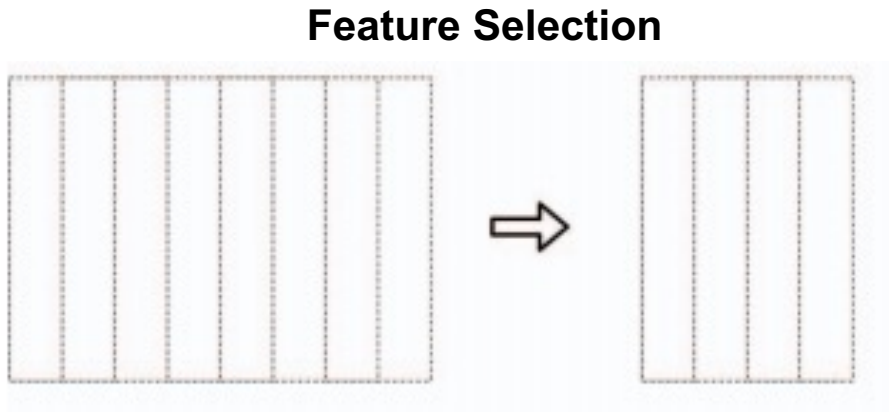


Maldición de la dimensionalidad

Feature selection

- Es el proceso de ordenar las variables por el valor de alguna función de puntuación.
- Para reducir la maldición de la dimensionalidad existen algunas estrategias:

- Discriminación manual, pero tiene limitaciones.
- Análisis correlacional de datos (Correlational Data Analysis, **CDA**)
- Análisis de componentes principales (Principal Component Analysis, **PCA**)



Maldición de la dimensionalidad

Ventajas de la reducción de dimensionalidad

- Ayuda en la reducción del espacio de almacenamiento.
- Reduce el tiempo de cálculo.
- Ayuda a eliminar variables redundantes, si las hay.

Desventaja de la reducción de dimensionalidad

- Si no se hace un análisis cuidadoso, puede provocar pérdida de datos valiosos.

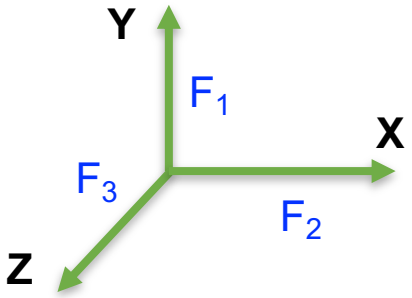
1. Análisis correlacional de datos

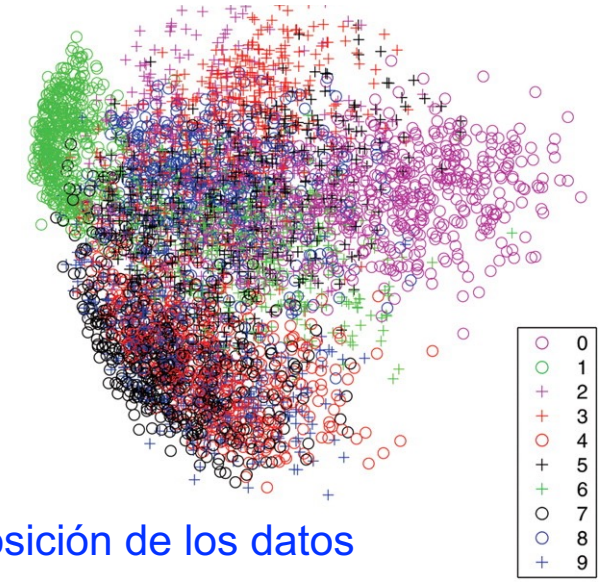
Análisis correlacional de datos

Correlaciones

- El **ACD (CDA)** es útil para reducir el número de variables, de un espacio de alta dimensión a uno de menor número de dimensiones.
- Esto se logra a través de la identificación de variables significativas.
- Esta identificación de correlaciones se utiliza para determinar el **grado de similitud** (relevancia/irrelevancia) de los valores de dos variables numéricas.
- Existe correlación entre 2 variables (**X**,**Y**) si al aumentar los valores de **X** también los hacen de **Y**, o viceversa.

	F_1	F_2	F_3
E_1	3	5	0
E_2	2	1	1
E_3	5	2	0
…	…	…	…
E_n	1	2	0





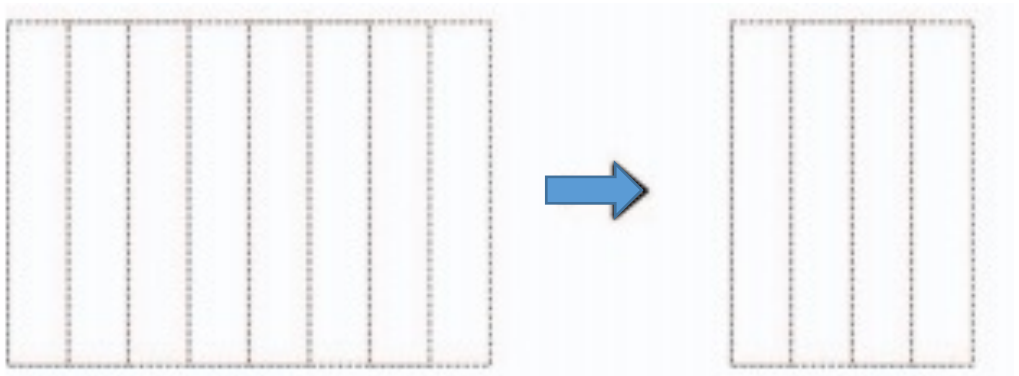
Existe una superposición de los datos

Análisis correlacional de datos

Correlaciones

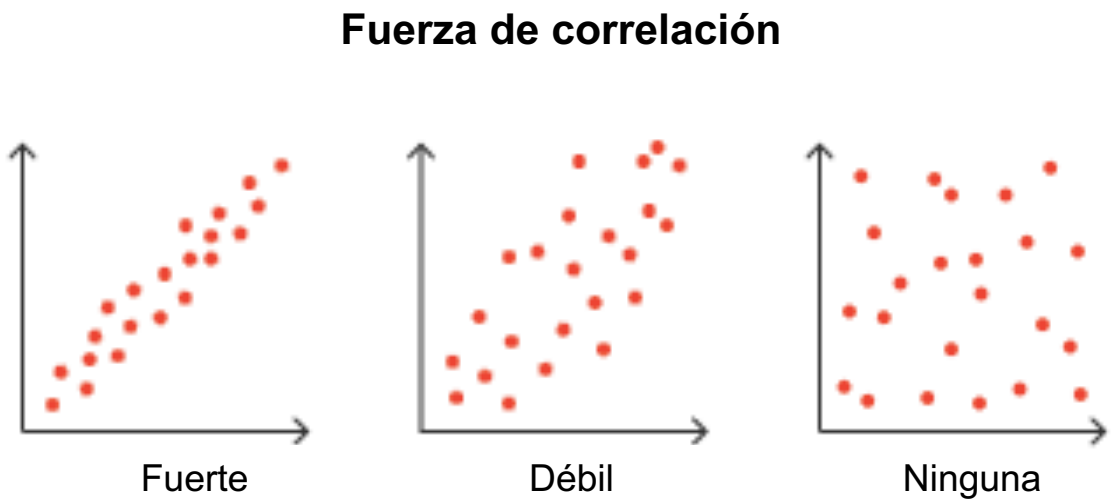
- La reducción consiste en que a partir de un conjunto de **variables originales:** $X_1, X_2, X_3, \ldots, X_n$
- Se obtiene otro subconjunto de **variables relevantes :** $X_1, X_2, X_3, \ldots, X_m$, donde $m < n$.

$(X_1, X_2, X_3, \ldots, X_n)$ $(X_1, X_2, X_3, \ldots, X_m)$

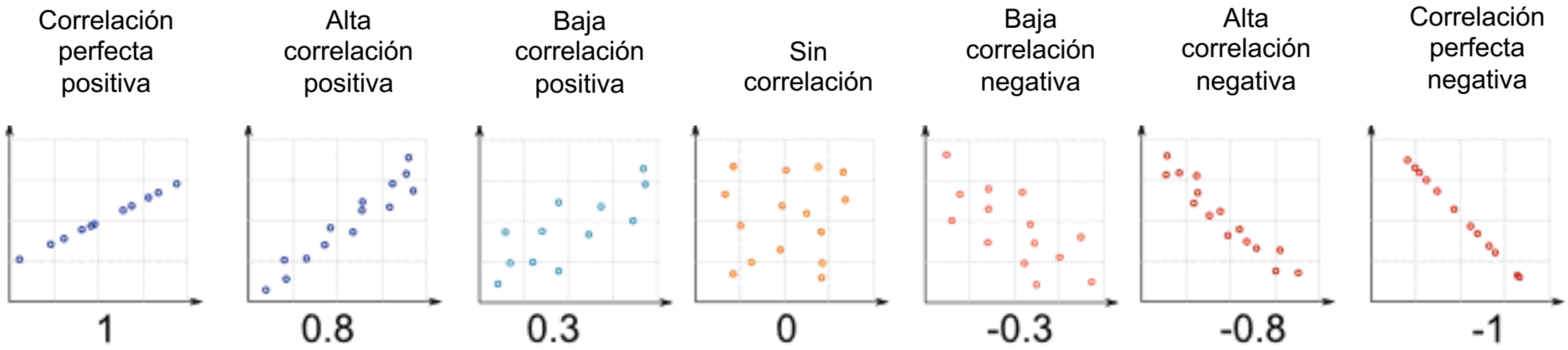


Evaluación visual de los datos

- Es importante hacer una evaluación visual de los datos a través de gráficos de dispersión (diagramas de dispersión).
- Estos gráficos utilizan una colección de puntos para mostrar los valores de dos variables.



Análisis correlacional de datos



Coeficiente de correlación

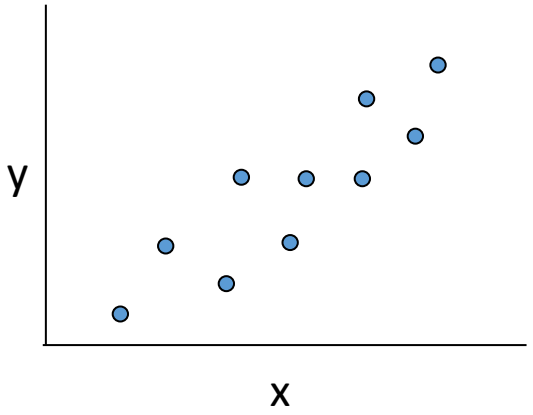
Los valores de correlación, conocidos como coeficiente de correlación de Pearson (su creador, Karl Pearson, 1857-1936), se define como:

Covarianza

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

$\bar{x}, \bar{y}$ *son las medias aritméticas de* x *e* y.

Varianza



Los valores de correlación, en este caso **r** o **R**, pueden variar entre -**1 y 1**.

Coeficiente de correlación

- Cuanto **más cerca está R de 1 o -1**, más fuerte es la correlación.
- Si **R es cercano a -1** las variables están correlacionadas negativamente.
- Si **R es cero** no existe correlación.

Intervalos utilizados para la identificación de correlaciones (Opción sugerida):

- De **-1.0 a -0.67** y **0.67 a 1.0** se conocen como correlaciones **fuertes o altas**.
- De **-0.66 a -0.34** y **0.34 a 0.66** se conocen como correlaciones **moderadas o medias**.
- De **-0.33 a 0.0** y **0.0 a 0.33** se conocen como correlaciones **débiles o bajas**.

Otras opciones utilizadas:

- De **-1.0 a -0.70** y **0.70 a 1.0** se conocen como correlaciones **fuertes o altas**.
- De **-0.69 a -0.31** y **0.31 a 0.70** se conocen como correlaciones **moderadas o medias**.
- De **-0.30 a 0.0** y **0.0 a 0.30** se conocen como correlaciones **débiles o bajas**.

Análisis correlacional de datos

Matriz de correlaciones

- Consiste en crear una **matriz** que aporta información sobre la relación entre pares de variables.
- El objetivo es obtener, a partir de esta matriz, un subconjunto de variables representativas que no tengan dependencia entre sí.

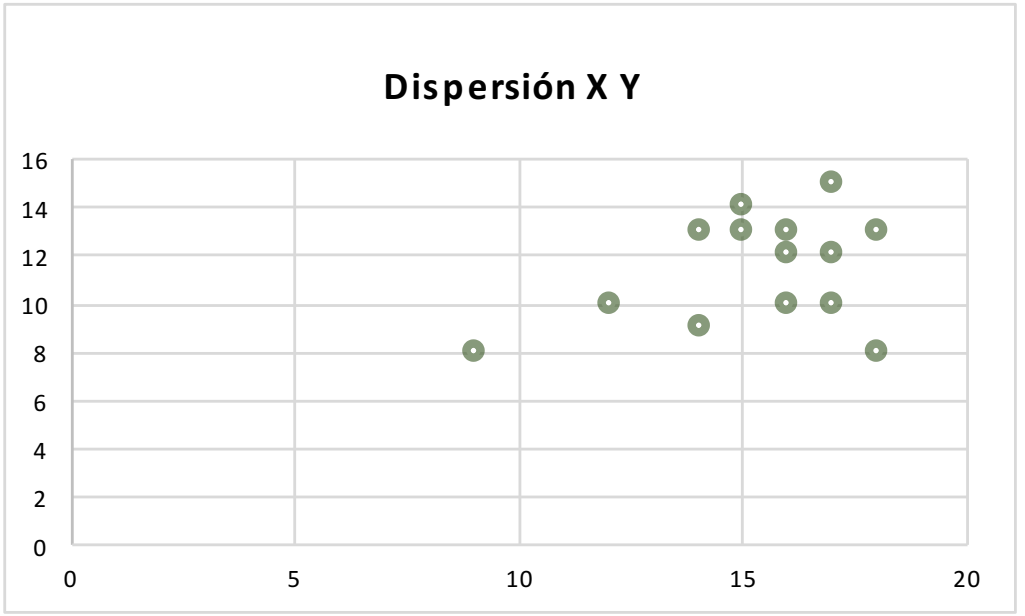
Ejemplo ilustrativo

Ejemplo ilustrativo

Sean las temperaturas de dos ciudades (**X** –Ciudad de México–, **Y** –Puebla–), determinar el coeficiente de correlación de Pearson:

Día	X	Y
Día 1	18	13
Día 2	17	15
Día 3	15	14
Día 4	16	13
Día 5	14	9
Día 6	12	10
Día 7	9	8
Día 8	15	13
Día 9	16	12
Día 10	14	13
Día 11	16	10
Día 12	18	8
Día 13	17	10
Día 14	17	12

Diagrama de dispersión



Ejemplo ilustrativo

Sean las temperaturas de dos ciudades (**X** –Ciudad de México–, **Y** –Puebla–), determinar el coeficiente de correlación de Pearson:

Día	X	Y	x = X-X'	y = Y-Y'	x2	y2	xy
Día 1	18	13	2.71	1.57	7.37	2.47	4.27
Día 2	17	15	1.71	3.57	2.94	12.76	6.12
Día 3	15	14	-0.29	2.57	0.08	6.61	-0.73
Día 4	16	13	0.71	1.57	0.51	2.47	1.12
Día 5	14	9	-1.29	-2.43	1.65	5.90	3.12
Día 6	12	10	-3.29	-1.43	10.80	2.04	4.69
Día 7	9	8	-6.29	-3.43	39.51	11.76	21.55
Día 8	15	13	-0.29	1.57	0.08	2.47	-0.45
Día 9	16	12	0.71	0.57	0.51	0.33	0.41
Día 10	14	13	-1.29	1.57	1.65	2.47	-2.02
Día 11	16	10	0.71	-1.43	0.51	2.04	-1.02
Día 12	18	8	2.71	-3.43	7.37	11.76	-9.31
Día 13	17	10	1.71	-1.43	2.94	2.04	-2.45
Día 14	17	12	1.71	0.57	2.94	0.33	0.98
Total	**214**	**160**			**78.86**	**65.43**	**26.29**
Media (X')	**15.29**						
Meda (Y')	**11.43**						

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

$$r = \frac{26.29}{\sqrt{78.86 * 65.43}} = \frac{26.29}{71.83} = \mathbf{0.36}$$

¿Qué pasa con variables cualitativas?

Variables cualitativas

En el caso de variables cualitativas

Pacientes, 7 variables:

- **Capacidad.** Capacidad del paciente para acudir a una consulta. (1-10)
- **Necesidad.** Importancia que le da el paciente a la consulta médica. (1-10)
- **Transporte.** Disponibilidad de transporte del paciente. (1-10)
- **Cuidado.** Disponibilidad para tener el cuidado de los niños. (1-10)
- **Permiso.** En caso de trabajar, facilidad para solicitar permisos médicos. (1-10)
- **Satisfacción.** Satisfacción del cliente con la atención médica. (1-10)
- **Facilidad.** Facilidad para obtener una cita y eficiencia de la misma. (1-10)
- **Visita.** Visita del paciente durante el último año (0 - no visitó, 1 - si visitó)

Variables cualitativas

En el caso de variables cualitativas

	Capacidad	Importancia	Transporte	Cuidado	Permiso	Satisfacción	Facilidad	Visita
Capacidad	1							
Importancia	-0.737	1						
Transporte	0.312	-0104	1					
Cuidado	0.312	-0104	0-379	1				
Permiso	0.277	0.060	0.623	0.623	1			
Satisfacción	0.220	-0.134	0.654	0.654	0.626	1		
Facilidad	0.389	-0.033	0.650	0.650	0.659	0.896	1	
Visita	0.396	-0.542	-0.503	-0.503	-0.425	-0.399	-0.328	1

- **R1.** Existe una relación fuerte (negativa) entre la **capacidad** que tiene el paciente para acudir a una consulta y la **Importancia** que le da el paciente a la consulta médica.
- **R2.** Se tiene una relación fuerte (positiva) entre la **satisfacción** del paciente con la atención médica y la **facilidad** que tiene para obtener una cita.