




# DATA ENGINEERING: DATOS Y PREPROCESAMIENTO



MSc Carlos Córdova  
BSc Carlos Ramírez



01

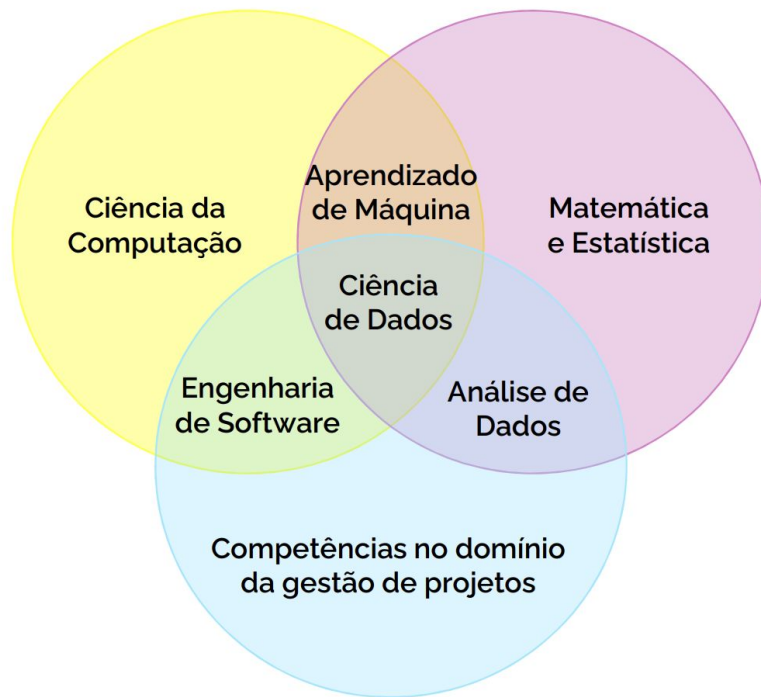
# Ciencia de Datos

¿En qué consiste la Ciencia de Datos?



## Ciencia de Datos

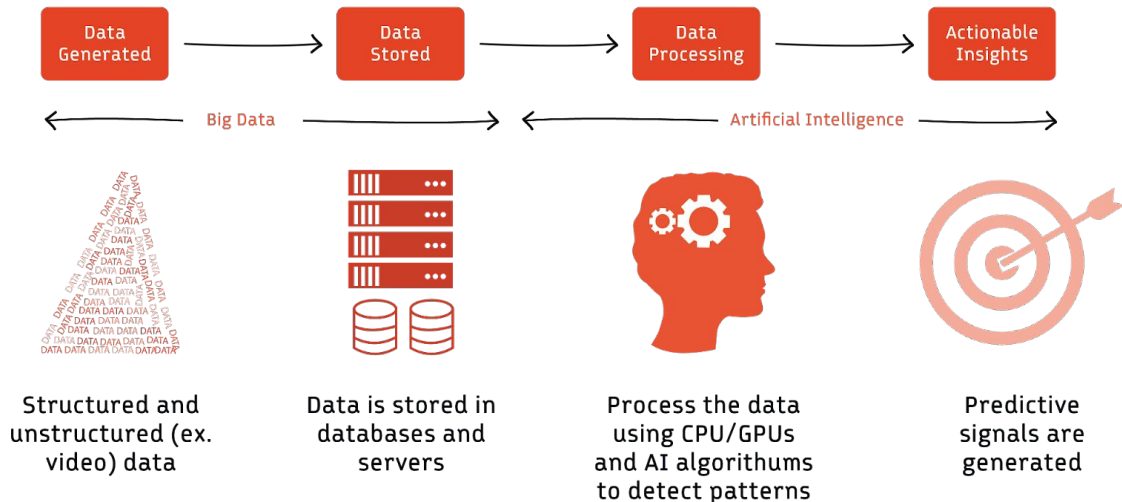
- Manejo y procesamiento de grandes volúmenes de datos
  - Estructurados
  - No estructurados
- Extracción de conocimiento
  - Toma de decisiones



## Ingeniería de Datos

- Construcción y mantenimiento de sistemas de datos.
- Extracción, **transformación y análisis de datos**.
- Asegurar accesibilidad y calidad de los datos.

### The Process



Central Processing Unit (CPU) Graphics Processing Unit (GPU)



02

# DATOS

Fuentes, tipos y estructura.



## Fuentes de datos

- Sensores:
  - Dispositivos IoT
  - Sensores ambientales
  - Equipos médicos
- Mediciones o recolección:
  - Encuestas y cuestionarios
  - Observaciones manuales
  - Registros históricos
- Simulaciones o computaciones:
  - Modelos predictivos
  - Simulaciones científicas

Los datos pueden ser:

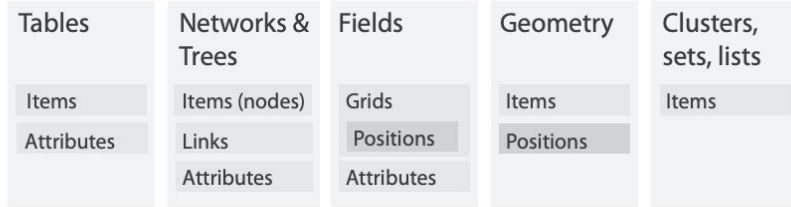
- Tratados
- Crudos (no tratados)

## Datasets

### → Data Types

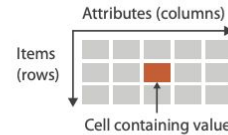
→ Items → Attributes → Links → Positions → Grids

### → Data and Dataset Types

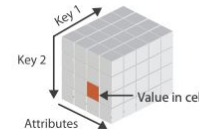


### → Dataset Types

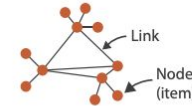
→ Tables



→ Multidimensional Table



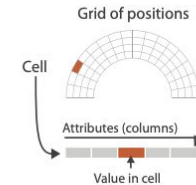
→ Networks



→ Trees



→ Fields (Continuous)



Lo que puede ser visualizado. T. Munzner Visualization Analysis & Design (Fig. 2.2)

## Estructura de un conjunto de datos

- **Dataset:**

- **n** instancias (filas, registros, elementos)
- **m** atributos (columnas, variables)

- **Tipos de atributos:**

- Valor único (e.g., número, cadena de texto, símbolo)
- Estructura compleja (listas, diccionarios, arrays)

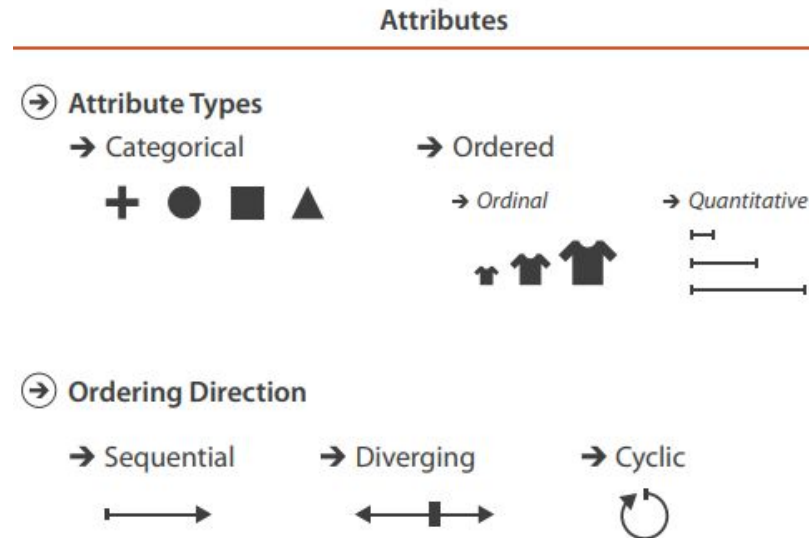
- **Tipos de variables:**

- Independiente
- Dependiente

	ID	Age	Sex	Weight	Height	Married	Migrantstatus
1	1	26	1	132	60	0	Nonmigrant
2	2	65	0	122	65	0	Migrant
3	3	15	1	184	67	0	Nonmigrant
4	4	7	1	145	59	0	Nonmigrant
5	5	80	0	100	64	0	Migrant
6	6	43	1	NA	NA	0	Nonmigrant
7	7	28	1	128	67	1	Nonmigrant
8	8	66	1	154	60	1	Nonmigrant
9	9	45	0	166	NA	0	Migrant
10	10	12	0	164	60	1	Migrant

## Tipos de Datos

- Clasificación:
  - **Ordenados:** numéricos, ordinales
  - **Categoricos:** nominales, no numéricos
- Atributos **ordenados:**
  - **Binarios:** 0 y 1
  - **Discretos:** valores enteros
  - **Continuos:** valores reales
- Atributos **categoricos:**
  - Nominales: valor de una **lista finita de posibilidades** (e.g., Sección A, B, C)
  - Ranqueados: valor categorico con **orden** (e.g., talla S, M, L, XL)
  - Arbitrarios: **lista infinita de opciones** sin orden (e.g., número de DNI)



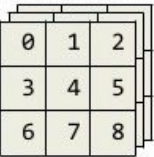
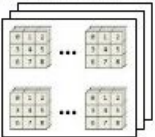


Tipos de atributos. T. Munzner, Visualization Analysis & Design (Fig. 2.7)



## Estructura de los datos

- Escalares:
  - Valor numérico individual
- Vectores:
  - Conjunto de escalares relacionados
  - Una dimensión adicional
- Tensores:
  - Generalización de escalares, vectores y matrices a más dimensiones
  - Arrays multidimensionales de datos

Dimensions	Example	Terminology
1		Vector
2		Matrix
3		3D Array (3 <sup>rd</sup> order Tensor)
N		ND Array



03

# PREPROCESAMIENTO DE DATOS

¿Qué hacer cuando los datos no están listos?

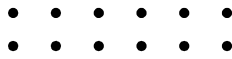


- 

• •

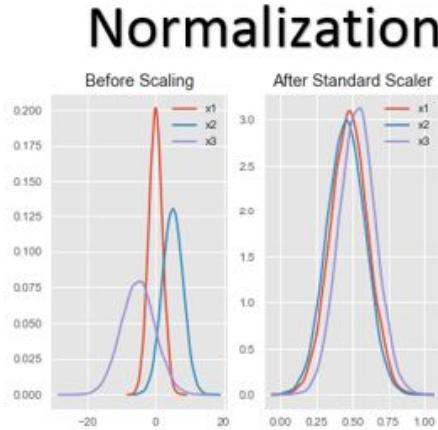
- Sistema de ventas:
  - Registros sin DNI asociado.
- Base de datos de pacientes:
  - Registros duplicados con variaciones en el nombre.
- Encuesta online:
  - Fechas de nacimiento en diferentes formatos.
  - Respuestas inconsistentes en mayúsculas y minúsculas.
- Registros de sensores:
  - Picos anómalos por errores de medición o interferencias



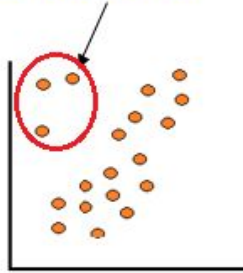


## El preprocesamiento de datos:

- Esencial para la data engineering.
- Asegura calidad y consistencia.



Outliers



Scaling

Imputation

	First	Second	Third
0	100.0	30.0	NaN
1	90.0	45.0	40.0
2	NaN	56.0	80.0
3	95.0	NaN	98.0

Encoding

Food Name	Apple	Chicken	Broccoli
Apple	1	0	0
Chicken	0	1	0
Broccoli	0	0	1

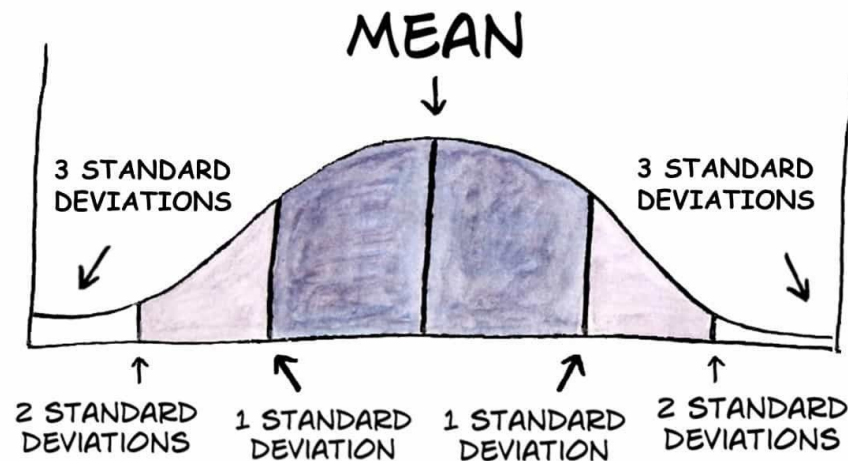
## Estadística

Permite:

- Detectar valores inválidos (e.g., errores de medidas de sensores)
- Identificar agrupamientos (e.g., medidas cercanas, similares entre instancias)
- Identificar atributos/variables redundantes (e.g., horas trabajadas, salario total)

## Medidas tradicionales

- Media
  - Agrega los datos y permite resumirlos
- Desviación estándar
  - Mide la dispersión de los datos
- Mediana, moda.



## Ejemplo: Detección de variables redundantes

- **Correlación**

- Medida estadística
- Permite medir la relación entre dos atributos/variables.

$x_i$  son los valores de una variable, por ejemplo, la edad.

$y_i$  son los valores de la otra variable, por ejemplo, el salario.

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

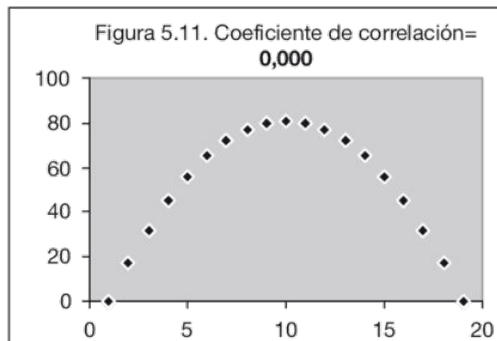
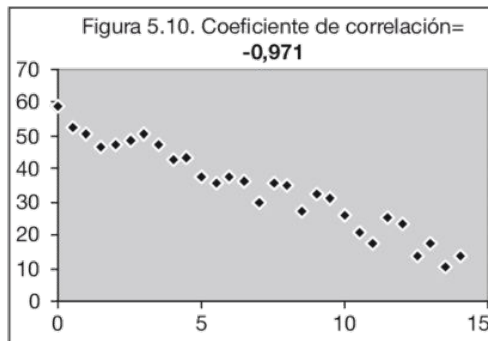
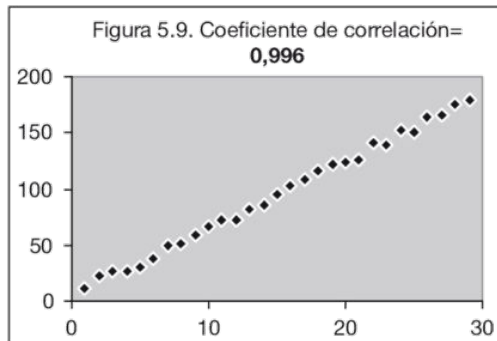
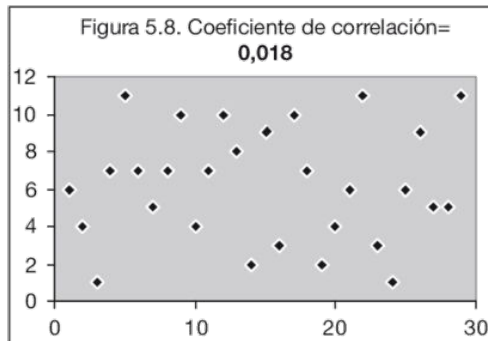
$r$  es el coeficiente de correlación de Pearson

$\bar{x}$  y  $\bar{y}$  son respectivamente los valores medios de las dos variables.

- Si la correlación entre dos variables es  $\pm 1$  o un valor próximo a  $\pm 1$  ambas se encuentran altamente correlacionadas.
  - Una debe ser removida.

## Ejemplo: Detección de variables redundantes

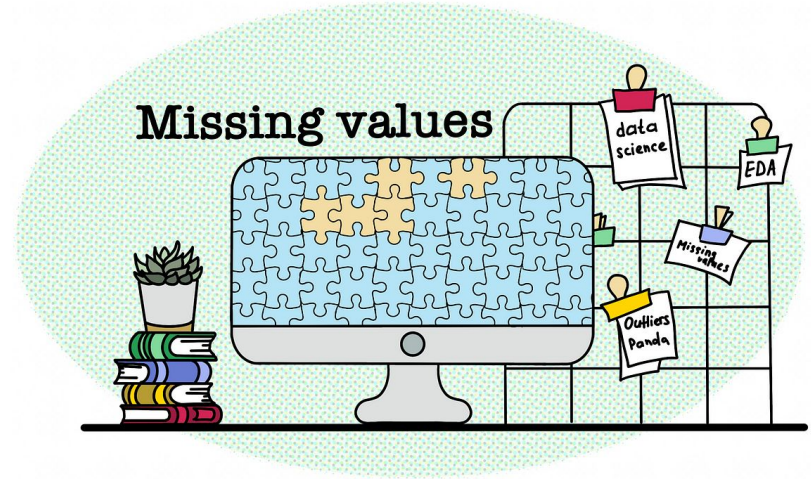
- Si la correlación entre dos variables es 1 o un valor próximo a 1, ambas se encuentran altamente correlacionadas.
  - Una debe ser removida.



- •
- • **Valores ausentes**
- •
- • Los datos ausentes afectan la calidad y precisión del análisis.
- •

Opciones:

- Descartar instancias incompletas:
  - Puede reducir significativamente el dataset.
- Agregar una bandera:
  - Permite descartar el atributo durante los cálculos (e.g., usar el 0).
- Imputar los datos:
  - Usar medidas clásicas como la media, mediana, moda.
  - Técnicas avanzadas:
    - KNN
    - Regresión Lineal





- • **Valores ausentes**
- • Los datos ausentes afectan la calidad y precisión del análisis.

Opciones:

- **Descartar instancias incompletas:**
  - Puede reducir significativamente el dataset.
- **Agregar una bandera:**
  - Permite descartar el atributo durante los cálculos (e.g., usar el 0).
- **Imputar los datos:**
  - Usar medidas clásicas como la media, mediana, moda.
  - Técnicas avanzadas:
    - KNN
    - Regresión Lineal

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	N/A	N/A	95%
8	Lite	76	77%
9	Fast+	180	N/A

← Delete

← Delete

← Delete



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
5	Lite	76	70%
6	Fast+	155	10%
8	Lite	76	77%

- •
- • **Valores ausentes**
- •
- • Los datos ausentes afectan la calidad y precisión del análisis.
- •

Opciones:

- Descartar instancias incompletas:
  - Puede reducir significativamente el dataset.
- **Agregar una bandera:**
  - Permite descartar el atributo durante los cálculos (e.g., usar el 0).


	col1	col2	col3	col4	col5			col1	col2	col3	col4	col5	
0	2	5.0	3.0	6	NaN	df.fillna(0)		0	2	5.0	3.0	6	0.0
1	9	NaN	9.0	0	7.0			1	9	0.0	9.0	0	7.0
2	19	17.0	NaN	9	NaN			2	19	17.0	0.0	9	0.0

- • **Valores ausentes**
- • Los datos ausentes afectan la calidad y precisión del análisis.

Opciones:

- Descartar instancias incompletas:
  - Puede reducir significativamente el dataset.
- Agregar una bandera:
  - Permite descartar el atributo durante los cálculos (e.g., usar el 0).
- **Imputar los datos:**
  - Usar medidas clásicas como la media, mediana, moda.
  - Técnicas avanzadas:
    - KNN
    - Regresión Lineal

Mode (Download Speed) = 200



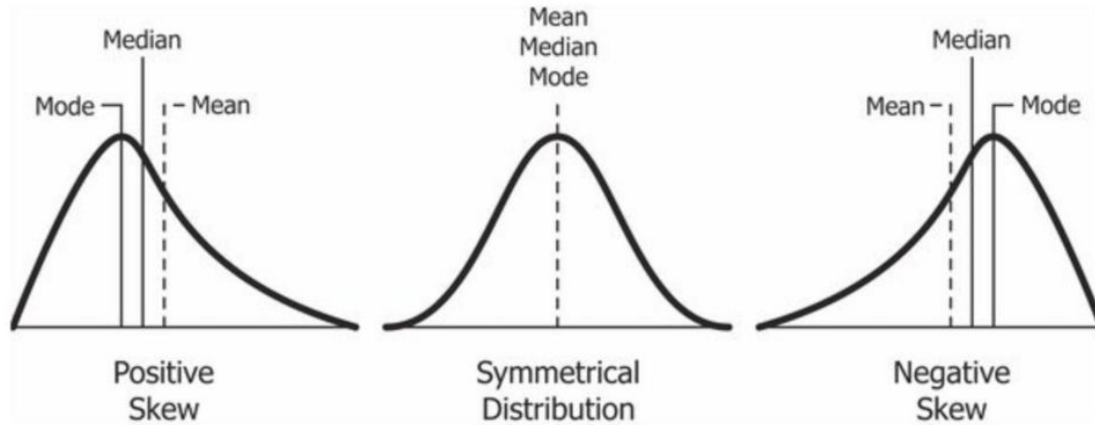
Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	N/A	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	N/A	95%
8	Lite	200	77%
9	Fast+	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	200	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	200	95%
8	Lite	200	77%
9	Fast+	180	95%

## Recomendación:

- Si los datos son categóricos, usar la **moda**.
- Si los datos son numéricos:
  - Si la distribución es normal: **media**.
  - Si la distribución no es normal (skewed):  
**mediana**.

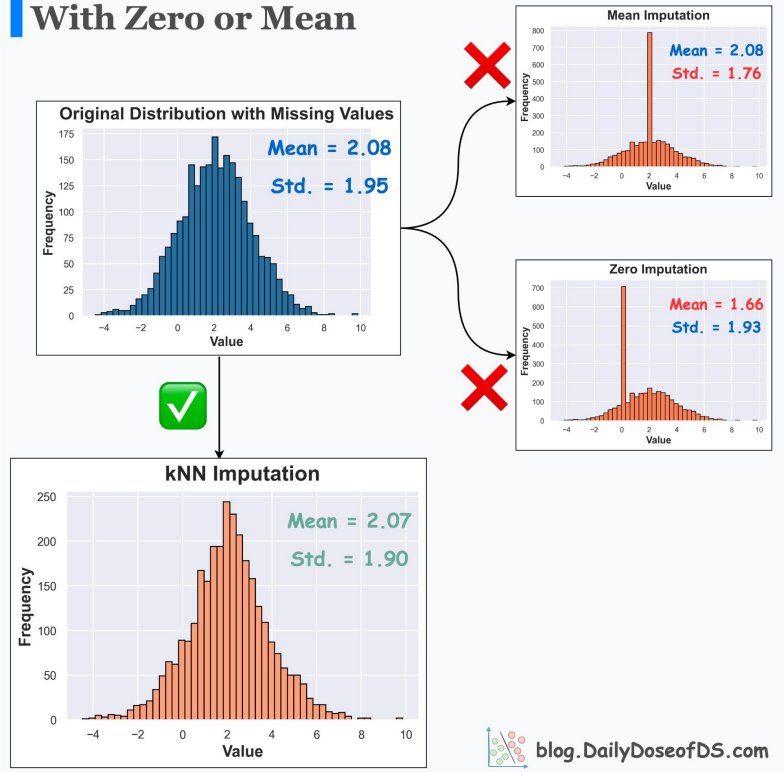


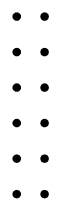
- • **Valores ausentes**
- • Los datos ausentes afectan la calidad y precisión del análisis.
- •

Opciones:

- Descartar instancias incompletas:
  - Puede reducir significativamente el dataset.
- Agregar una bandera:
  - Permite descartar el atributo durante los cálculos (e.g., usar el 0).
- **Imputar los datos:**
  - Usar medidas clásicas como la media, mediana, moda.
  - Técnicas avanzadas:
    - **KNN**
    - Regresión Lineal

## Avoid Filling Missing Values With Zero or Mean





## Normalización

Distintos atributos pueden encontrarse en escalas diferentes.

- Resultados distorsionados.
- Comparaciones tendenciosas.
  - Favorecimiento a atributos con mayor escala.

:Summary Statistics:

	Min	Max	Mean	SD	Class Correlation	
sepal length:	4.3	7.9	5.84	0.83	0.7826	
sepal width:	2.0	4.4	3.05	0.43	-0.4194	
petal length:	1.0	6.9	3.76	1.76	0.9490	(high!)
petal width:	0.1	2.5	1.20	0.76	0.9565	(high!)

## Normalización

La normalización:

- Elimina la influencia de la escala.
- Transforma los valores de la escala entre 0 y 1.

The diagram shows the normalization formula  $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$  with arrows pointing from descriptive labels to each component of the formula.

Normalized Value  $\rightarrow x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

Original Value  $\rightarrow x$

Maximum Value of  $x \rightarrow \max(x)$

Minimum Value of  $x \rightarrow \min(x)$

## Normalización

La normalización:

- Elimina la influencia de la escala.
- Transforma los valores de la escala entre 0 y 1.
- **No mantiene la dispersión de los datos.**

Alternativa: **Estandarización (Standardization)**

- Transforma los valores para que la media sea 0 y la desviación estándar sea 1.
- Mantiene la dispersión de los datos.

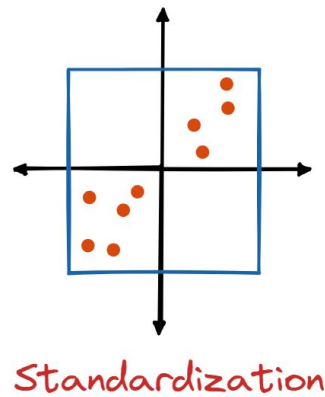
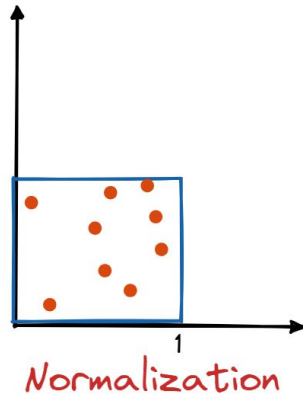
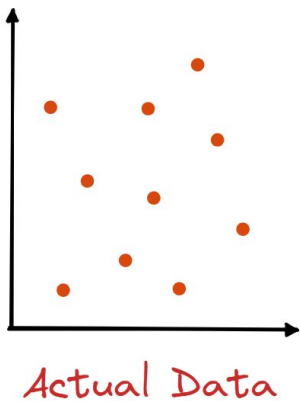
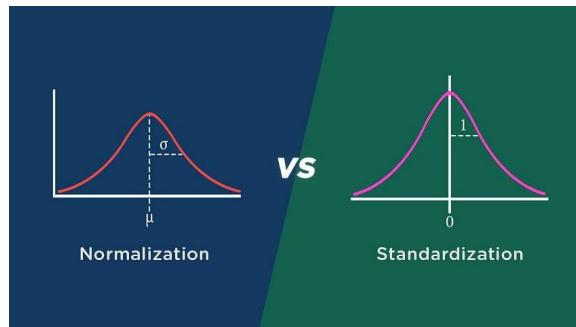
$$X_{scaled} = \frac{X - X_{mean}}{X_{stddev}}$$



# Normalización

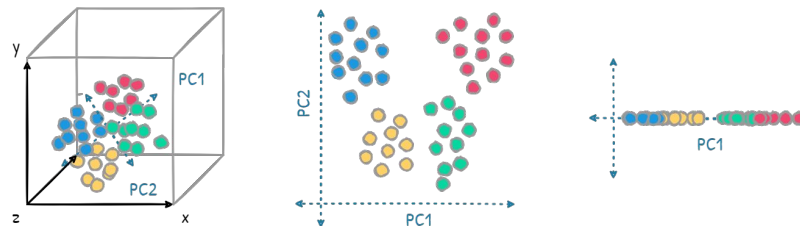
## Alternativa: Estandarización (Standardization)

- Transforma los valores para que la media sea 0 y la desviación estándar sea 1.
- Mantiene la dispersión de los datos.
- Desventaja: no acota los valores a un rango específico.



## Reducción de dimensionalidad

- El exceso de variables puede dificultar el análisis.
- Las variables correlacionadas pueden generar redundancia.
- La reducción de dimensionalidad acentúa y facilita el filtrado del ruido (outliers).



A diagram consisting of two vertical columns of six dots each, representing a 6x2 grid.

• •

- 



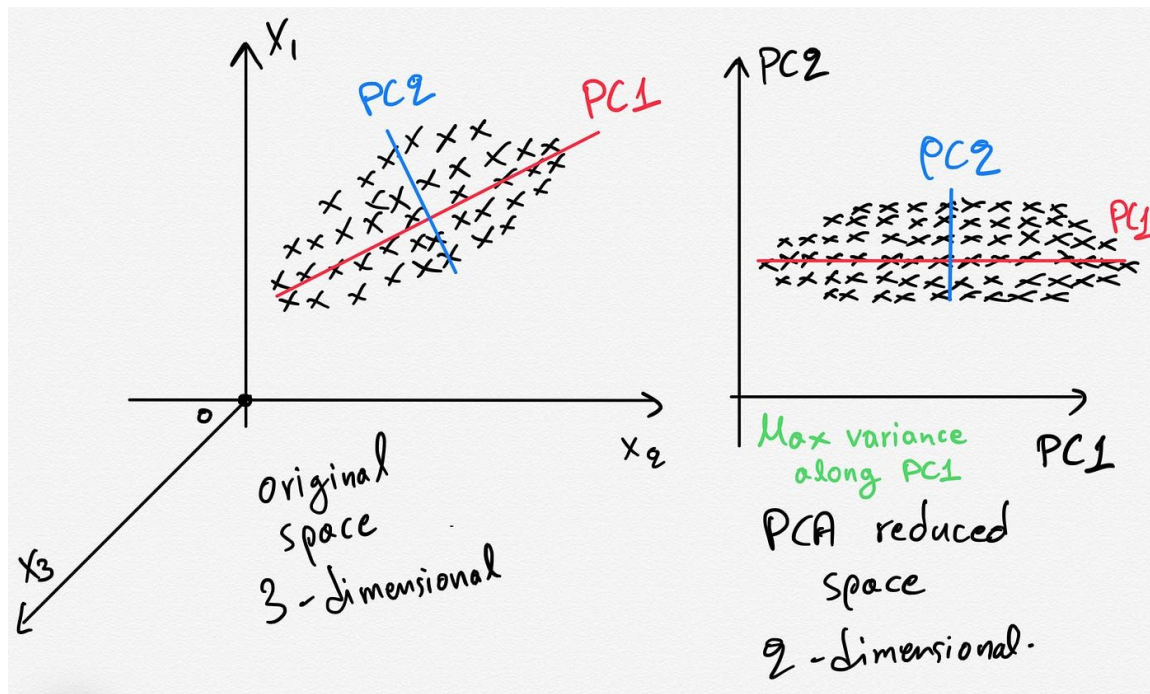
## Reducción de dimensionalidad

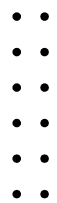
- **PCA (Análisis de Componentes Principales)**

- [Vídeo explicativo](#)
- Resumido:
  - Se estandarizan los datos.
  - Se calcula una matriz de covarianza para determinar cómo varían los pares de variables.
  - Se calculan los valores y vectores propios (eigenvalues y eigenvectors)
  - Se seleccionan los vectores propios que explican la mayor parte de la varianza.
  - Se proyectan los datos originales en el nuevo espacio de componentes principales.

## Reducción de dimensionalidad

- PCA (Análisis de Componentes Principales)





## **Mapeo de datos nominales a números**

Los algoritmos de ML trabajan sobre números.

Opciones:

- Para datos nominales ranqueados (con orden):
  - El mapeo es directo (e.g., pequeño → 1, mediano → 2, grande → 3).
  - El orden se mantiene.
- Para datos nominales no ranqueados:



## • • Mapeo de datos nominales a números

- Los algoritmos de ML trabajan sobre números.

Opciones:

- Para datos nominales no ranqueados:
  - **Label Encoding**
    - Asigna un valor único a cada categoría
    - Simple de realizar, pero introduce un orden implícito

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29



Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

## • • Mapeo de datos nominales a números

- • Los algoritmos de ML trabajan sobre números.

Opciones:

- Para datos nominales no ranqueados:
  - **One-Hot Encoding**
    - Asigna una columna booleana por cada valor único.
    - Aumenta la dimensionalidad.

Original Data		One-Hot Encoded Data			
Team	Points	Team_A	Team_B	Team_C	Points
A	25	1	0	0	25
A	12	1	0	0	12
B	15	0	1	0	15
B	14	0	1	0	14
B	19	0	1	0	19
B	23	0	1	0	23
C	25	0	0	1	25
C	29	0	0	1	29

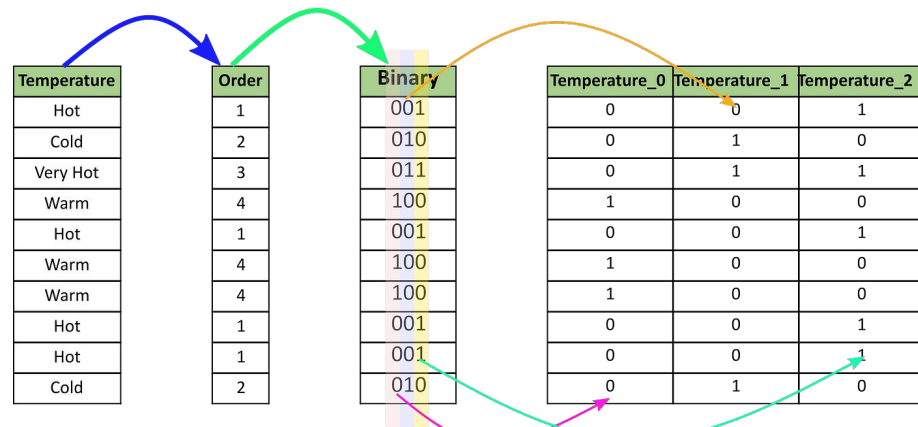


## • • Mapeo de datos nominales a números

- Los algoritmos de ML trabajan sobre números.

Opciones:

- Para datos nominales no ranqueados:
  - **Binary Encoding**
    - Transforma cada categoría única a un número binario.
    - Genera las columnas necesarias para formar todos los números binarios hasta el total de categorías únicas.
    - Reduce la dimensionalidad pero es menos intuitivo.





# ¿Preguntas?

