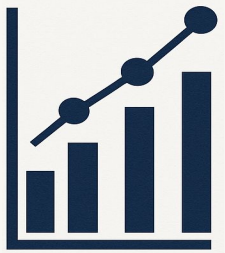

# PREPARACIÓN DE DATOS

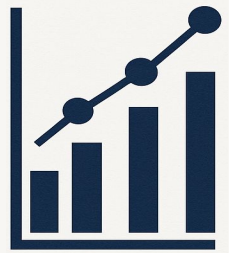
Integración - Limpieza - Transformación

# ¿QUÉ VEREMOS HOY?



1. Transformación
    a. Discretización
    b. Numerización
    c. Normalización y estandarización
    d. Otros métodos

# TRANSFORMACIÓN



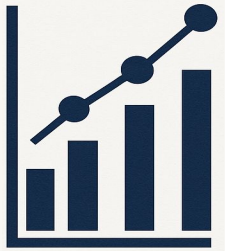
## DISCRETIZACIÓN

Conversión de un valor numérico a un valor categórico ordinal

| Nota | Valor | categoría |
|---|---|---|
| [0;4) | 1 | No Satisfactorio |
| [4;6) | 2 | Satisfactorio |
| [6;8) | 3 | Bueno |
| [8;9) | 4 | Muy Bueno |
| [9;10] | 5 | Excelente |

| Compra Super | Valor | categoría |
|---|---|---|
| [0;5) | 1 | Muy Bajo |
| [5;10) | 2 | Bajo |
| [10;20) | 3 | Medio |
| [20;50) | 4 | Alto |
| [50;+inf] | 5 | Muy Alto |

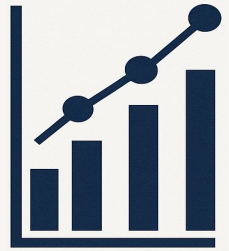
# TRANSFORMACIÓN



## DISCRETIZACIÓN

- **Uniforme:** Cada contenedor tiene el mismo ancho en el lapso de valores posibles para la variable.
- **Cuantil:** Cada contenedor tiene la misma cantidad de valores, divididos según percentiles.
- **Agrupado:** Se identifican grupos y se asignan ejemplos a cada grupo.

# TRANSFORMACIÓN



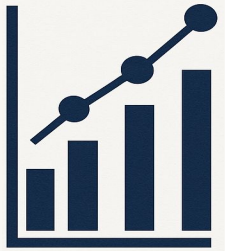
## NUMERIZACIÓN

Conversión de un valor categórico/nominal a un valor numérico

| Categoría | Valor |
|---|---|
| Sin estudios | 1 |
| Primario | 2 |
| Secundario | 3 |
| Terciario | 4 |
| Universitario | 5 |

| | Sin estudios | Primario | Secundario | Terciario | Universitario |
|---|---|---|---|---|---|
| Primario | 0 | 1 | 0 | 0 | 0 |
| Secundario | 0 | 0 | 1 | 0 | 0 |
| Terciario | 0 | 0 | 0 | 1 | 0 |
| Universitario | 0 | 0 | 0 | 0 | 1 |

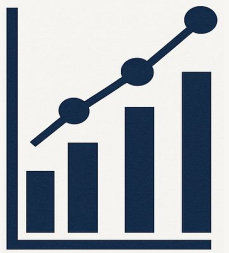
# TRANSFORMACIÓN



## NUMERIZACIÓN

- **Numerización 1 a n:** Si una variable nominal x tiene n posibles valores, entonces creamos n variables numéricas, con valores {0,1} dependiendo de si la variable nominal toma ese valor o no.
- **Numerización 1 a 1:** Se identifica un cierto orden o magnitud en los valores de un atributo nominal.

# TRANSFORMACIÓN



## NORMALIZACIÓN - ESTANDARIZACIÓN

Existen técnicas de análisis de datos que requieren que los datos se encuentren normalizados en un mismo rango.

# TRANSFORMACIÓN

## NORMALIZACIÓN - ESTANDARIZACIÓN

1) Normalización Lineal Uniforme

$$x' = \frac{x - \min}{\max - \min}$$

# TRANSFORMACIÓN

## NORMALIZACIÓN - ESTANDARIZACIÓN

2) Estandarización

$$x' = \frac{x - \mu}{\sigma}$$

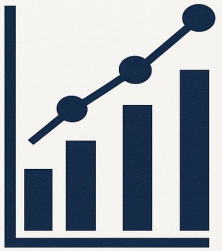
# TRANSFORMACIÓN

NORMALIZACIÓN - ESTANDARIZACIÓN

3) Escalado sigmoidal (softmax)

$$x' = \frac{1}{1 + e^{-x}}$$

$$x' = \frac{1}{1 + e^{-20(x-0.5)}}$$

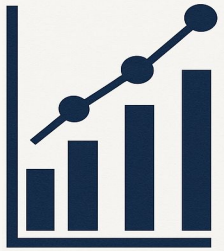


# TRANSFORMACIÓN

## NORMALIZACIÓN - ESTANDARIZACIÓN

| Característica | Normalización | Estandarización |
|---|---|---|
| **Qué hace** | Escala los datos a un **rango fijo** (ej. [0,1]) | Ajusta los datos a una distribución con **media = 0** y **desvío estándar = 1** |
| **Fórmula** | $x' = \frac{x - \min}{\max - \min}$ | $x' = \frac{x - \mu}{\sigma}$ |
| **Resultado** | Todos los valores quedan entre 0 y 1 (a veces -1 y 1) | Valores convertidos a *z-scores* (pueden ser negativos o mayores a 1) |
| **Ejemplo** | Notas de 0 a 10 → Nota 7 → (7-0)/(10-0)= **0.7** | Notas con media=6.5 y σ=1.8 → Nota 7 → (7-6.5)/1.8 = **0.28** |
| **Cuándo usar** | - Cuando los datos tienen **rangos muy diferentes**<br>- Algoritmos basados en distancias (**KNN, Redes Neuronales, Clustering**) | - Cuando te importa la **distribución estadística**<br>- Algoritmos que suponen normalidad (**Regresión, SVM, PCA**) |
| **Ventaja** | Pone todo en la misma escala fácil de interpretar | Permite comparar variables en función de la desviación respecto a la media |
| **Desventaja** | Muy sensible a **outliers** (un valor extremo deforma la escala) | Los valores no quedan en un rango fijo, pueden ser muy grandes/pequeños |

# TRANSFORMACIÓN



## OTROS MÉTODOS

- Derivación: obtiene un nuevo atributo a partir de realizar cálculos con valores de otros atributos. (ingreso_per_cápita, índice de masa muscular)
- Interacciones: representa el producto, la razón o la diferencia entre columnas.
- Variables temporales: formatea fechas usando los datos día, mes, año. También se puede calcular la diferencia temporal.
- Y más . . .