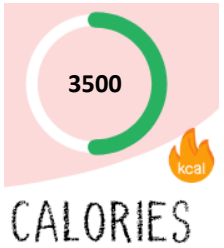
# Escalamiento de variables

# Qué es?

Transformar datos **numéricos** en una escala estándar…. Por ejemplo entre 0 y 1.



# Importante

Se escala por columna (variable), no por fila

| Pais | Continente | Mes | Expectativa de vida | # Habitantes |
|---|---|---|---|---|
| Afganisthan | Asia | Enero | 28.80 | 8425333 |
| Afganisthan | Asia | Febrero | 31.332 | 9240934 |
| Afganisthan | Asia | Marzo | 31.997 | 10267083 |
| Afganisthan | Asia | Abril | 34.020 | 11537966 |
| Afganisthan | Asia | Mayo | 36.088 | 13079460 |

# Tipos communes de escalamiento

## Normalización:

- Valores entre 0 y 1
- Más sensible a valores atípicos
- No modifica la distribución original
- Útil cuando la distribución de tus datos no es normal

### Cuándo utilizarla?

- Cuando tienes algoritmos que se basan en magnitud o distancia (KNN, Redes Neuronales, SVMs con kernel RBF)

## Estandarización:

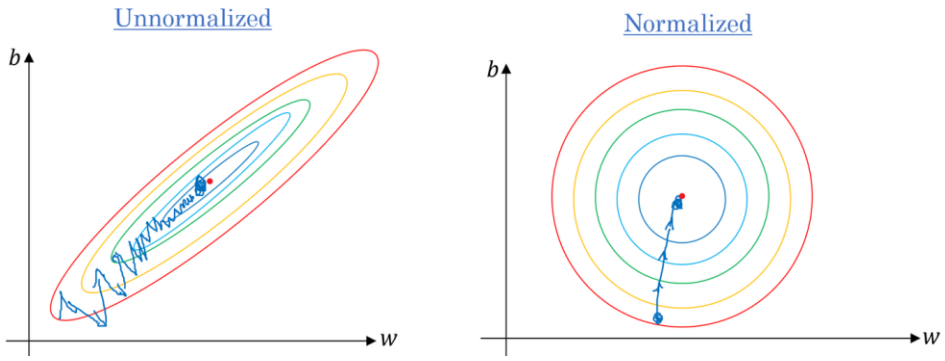
- Transforma los datos para que tengan media 0 y std 1
- Menos afectada por los valores atípicos
- Útil cuando los datos están normalmente distribuidos

### Cuándo utilizarla?

- Cuando tienes algoritmos que asumen distribuciones normales (Regresión Lineal, Regresión Logística, PCA)

# Porqué escalar los datos?

- Métodos que utilizan **gradiente descendente** pueden batallar en converger sin el escalamiento



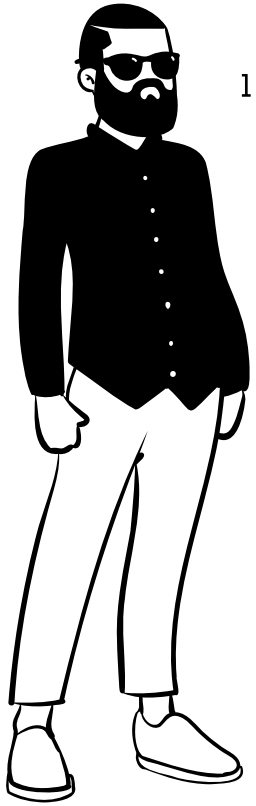

- Algunos modelos se verán sesgados por diferentes escalas, ya que se basan en **métricas de distancia**.

# Quién es más similar?

A

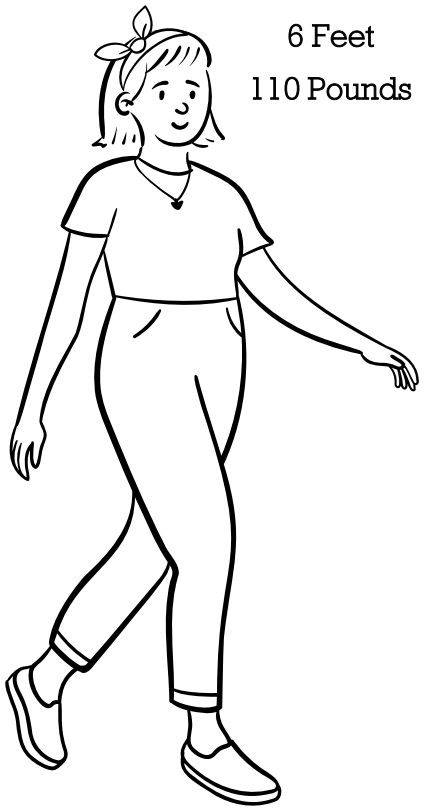
6 Feet
187 Pounds

B



6 Feet
175 Pounds

C



6 Feet
110 Pounds

D

5 Feet
100 Pounds