Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

CIMAT

Motivación: ¿por qué escalar/normalizar antes de modelar?

- Muchos métodos se basan en **distancias** o **varianzas**. Si las magnitudes difieren (p. ej., ingresos en miles vs. edad en años), unas variables *dominan* el análisis.[1]
- En **PCA**, sin estandarizar, los componentes quedan sesgados hacia variables de gran rango; en **k-medias**, la métrica Euclídea queda desbalanceada.
- **Visualización exploratoria** (EDA) no es decorativa: diagnostica asimetrías, outliers y escalas; guía la elección de transformación.[2]

[1] Hastie et al. (2009), cap. 3.

[2] Tukey (1977); Cleveland (1993); Wilkinson (2005).

Panorama: enfoques clásicos y robustos (visión general)

Clásicos

- **Min–Max**: reescala a $[0,1]$. Sensible a outliers.
- **Z-score**: $z = (x - \mu)/\sigma$. Útil para PCA, métodos basados en distancia y regularización.

Robustos

- **Estandarización robusta**: $(x - \mathrm{mediana})/\mathrm{RIQ}$, atenúa el efecto de outliers.
- **Transformaciones**: log, Box–Cox, Yeo–Johnson, cuando hay asimetrías fuertes o efectos multiplicativos.

Referencias: Rousseeuw & Leroy (1987); Hastie et al. (2009); James et al. (2021).

Buenas prácticas y riesgos comunes

Evitar *data leakage*

Ajustar el escalador **sólo con entrenamiento** y aplicar luego a validación/prueba con esos parámetros.[a]

[a] Kuhn & Johnson (2013), cap. 4.

Reproducibilidad

Encadenar pasos en un **pipeline**: imputación → codificación → escalamiento → modelo, con `random_state` y semillas documentadas.

Diagnóstico previo vía EDA

Decidir *antes* de modelar: ¿asimetrías fuertes? ⇒ transformación; ¿outliers? ⇒ esquema robusto. La EDA está en la ruta de tu Unidad 1.[a]

[a] Flujo de proyecto con preprocesamiento.

Normalización (Min–Max)

Definición

Dado un vector de datos $x_1, \ldots, x_n$, la transformación min–max es:

$$x_i^* = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad \in [0, 1].$$

- Escala todos los valores al rango $[0, 1]$.
- Útil cuando las variables tienen límites naturales (p. ej., proporciones, imágenes).
- Muy sensible a **outliers**, ya que dependen de extremos.

Referencias: Bishop (2006); James et al. (2021, cap. 3).

Propiedades del escalamiento Min–Max

- **Invarianza afín:** Si $x_i^* = (x_i - \min(x))/(\max(x) - \min(x))$, entonces:

$$y_i = ax_i + b \quad \Rightarrow \quad y_i^* = x_i^*, \quad a > 0.$$

  Es decir, es invariante a traslaciones y escalas positivas.

- **Rango controlado:**

$$\min(x^*) = 0, \quad \max(x^*) = 1.$$

- **Monotonía:**

$$x_i < x_j \quad \Rightarrow \quad x_i^* < x_j^*.$$

  Preserva el orden de los datos.

- **Sensibilidad:** Un solo valor extremo puede modificar drásticamente todos los x_i^*.

Referencias: Bishop (2006, cap. 2); Hastie et al. (2009, cap. 3).

Estandarización (Z-score)

Definición

Sea $x_1, \ldots, x_n$ con media $\bar{x}$ y desviación estándar s, el Z-score es:

$$z_i = \frac{x_i - \bar{x}}{s}, \qquad \text{con media } 0 \text{ y varianza } 1.$$

- Centra los datos en torno a cero y los hace adimensionales.
- Es la base para métodos que dependen de varianzas y covarianzas (p. ej., PCA).
- Permite interpretar coeficientes estandarizados en regresión (comparar la magnitud de los efectos).

Referencias: Jolliffe & Cadima (2016, PCA); Hastie et al. (2009).

Propiedades de la estandarización (Z-score)

- **Media y varianza:**

$$\bar{z} = 0, \quad \mathrm{Var}(z) = 1.$$

- **Invarianza a traslaciones:**

$$(x_i + c) \mapsto z_i \quad \Rightarrow \quad \text{mismo resultado.}$$

- **Invarianza a cambios de escala positiva:**

$$(ax_i),\ a > 0 \quad \mapsto \quad z_i \quad \text{no cambia.}$$

- A diferencia de Min–Max, los z_i pueden ser negativos y no quedan acotados.
- La definición requiere $\bar{x}$ y s, lo que la hace sensible a outliers.

Referencias: Jolliffe & Cadima (2016); Hastie et al. (2009).

Comparación: Normalización vs Estandarización

Normalización (Min–Max)

- Rango fijo $[0, 1]$.
- Útil en redes neuronales e imágenes.
- Sensible a valores extremos.

Estandarización (Z-score)

- Media 0, varianza 1.
- Útil en PCA, regresión penalizada, k-means.
- Menos sensible a cambios de escala, pero aún afectada por outliers.

¿Cuál usar? Depende del modelo, el tipo de variable y la presencia de outliers.

¿Por qué escaladores robustos?

- Tanto Min–Max como Z-score son sensibles a **outliers**.
- Un solo valor extremo puede alterar:
  - El rango completo (en Min–Max).
  - La media y la desviación estándar (en Z-score).
- Los **escaladores robustos** usan estadísticas resistentes (mediana, rango intercuartílico).
- Objetivo: reducir la influencia de observaciones atípicas manteniendo comparabilidad entre variables.

Referencia: Rousseeuw & Leroy (1987).

Escalador robusto: Mediana y RIQ

Definición

Dado un vector $x_1, \ldots, x_n$ con mediana m y rango intercuartílico $RIQ = Q_{0.75} - Q_{0.25}$:

$$x_i^R = \frac{x_i - m}{RIQ}.$$

- Centra los datos en la mediana y los escala según la dispersión central.
- **Robusto**: mediana y cuartiles tienen un *breakdown point* del 50%.
- Rango típico de valores: $[-1.34, 1.34]$ contiene aproximadamente 50% de los datos en una Normal estándar.

Referencia: Huber & Ronchetti (2009).

Propiedades del escalador robusto

- **Invarianza al orden**: preserva la monotonía de los datos.
- **Resistencia a outliers**: un valor extremo no modifica drásticamente m ni RIQ.
- **No acotado**: a diferencia del Min–Max, x_i^R puede tomar valores arbitrariamente grandes.
- **Interpretación**: mide cuántos rangos intercuartílicos está alejado un valor de la mediana.

Referencias: Rousseeuw & Leroy (1987); Huber & Ronchetti (2009).

Transformaciones no lineales para robustez

- **Logaritmo**:
$$y = \log(x+1) \quad \text{(cuando } x \geq 0\text{)}.$$
Reduce asimetrías y comprime colas largas.
- **Box–Cox** (Box & Cox, 1964):
$$y(\lambda) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(x), & \lambda = 0. \end{cases}$$
- **Yeo–Johnson** (1999): extensión que admite $x \in \mathbb{R}$.

Referencias: Box & Cox (1964); Yeo & Johnson (2000).

Nota

En aplicaciones prácticas se suele usar $\log(1+x)$ para admitir ceros en los datos.

Propiedades de Box–Cox

- Parametriza una familia de transformaciones continuas en λ.
- **Objetivo**: aproximar normalidad y estabilizar varianza.
- La elección de λ se realiza por máxima verosimilitud bajo el supuesto Normal:

$$\hat{\lambda} = \arg\max_{\lambda} \ell(\lambda; x).$$

- **Limitación**: requiere $x > 0$.

Referencia: Box & Cox (1964).

Transformación de Yeo–Johnson

Definición

$$y(\lambda) = \begin{cases} \dfrac{(x+1)^\lambda - 1}{\lambda}, & x \geq 0,\ \lambda \neq 0, \\ \log(x+1), & x \geq 0,\ \lambda = 0, \\ -\dfrac{(-x+1)^{2-\lambda} - 1}{2-\lambda}, & x < 0,\ \lambda \neq 2, \\ -\log(-x+1), & x < 0,\ \lambda = 2. \end{cases}$$

- Admite valores negativos de x (a diferencia de Box–Cox).
- Útil en contextos con variables centradas o simétricas alrededor de 0.

Referencia: Yeo & Johnson (2000).

Referencias clave

- T. Hastie, R. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
- G. James, D. Witten, T. Hastie, R. Tibshirani (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer.
- J. W. Tukey (1977). *Exploratory Data Analysis*. Addison–Wesley.
- W. S. Cleveland (1993). *Visualizing Data*. Hobart Press.
- L. Wilkinson (2005). *The Grammar of Graphics* (2nd ed.). Springer.
- P. J. Rousseeuw, A. M. Leroy (1987). *Robust Regression and Outlier Detection*. Wiley.
- M. Kuhn, K. Johnson (2013). *Applied Predictive Modeling*. Springer.

Referencias de esta sección

- P. J. Rousseeuw, A. M. Leroy (1987). *Robust Regression and Outlier Detection*. Wiley.
- P. J. Huber, E. M. Ronchetti (2009). *Robust Statistics* (2nd ed.). Wiley.
- G. E. P. Box, D. R. Cox (1964). "An Analysis of Transformations". *Journal of the Royal Statistical Society, Series B*, 26(2), 211–252.
- I. K. Yeo, R. A. Johnson (2000). "A New Family of Power Transformations to Improve Normality or Symmetry". *Biometrika*, 87(4), 954–959.
- T. Hastie, R. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
- I. T. Jolliffe, J. Cadima (2016). "Principal component analysis: a review and recent developments". *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.