# ¿Cuándo Normalizar y cuándo Estandarizar?

El preprocesamiento de datos es una etapa que puede influir directamente en el rendimiento de un modelo. Uno de los pasos más importantes dentro de esta fase es **el escalado de características**.

Cuando las variables tienen rangos muy diferentes, **algunos algoritmos pueden verse afectados y entender que aquellos con valores más grandes tienen mayor importancia, aunque estos no sean necesariamente más relevantes para nuestra predicción**.

⚙ **La Normalización consiste en transformar los datos para que se encuentren en un rango específico, generalmente entre 0 y 1**. Este método es especialmente útil cuando los datos no siguen una distribución normal y tienen valores con un rango bien definido. Los algoritmos basados en distancia, como **K-Nearest Neighbors (KNN) o los métodos de clustering**, dependen de medidas comparativas entre observaciones, por lo que necesitan que todas las variables estén en escalas similares para evitar sesgos.

⚙ **La Estandarización ajusta los datos para que tengan una media de 0 y una desviación estándar de 1**.
A diferencia de la normalización, **no limita los valores a un rango fijo, sino que modifica la distribución de los datos sin alterar su forma relativa**. Esta técnica es ideal cuando las variables siguen una distribución normal o aproximadamente normal, ya que muchos algoritmos de machine learning, como la **regresión logística o las máquinas de soporte vectorial (SVM)**, asumen que los datos están distribuidos de esta manera.

Además, **la estandarización es más resistente a valores extremos** en comparación con la normalización, lo que la hace más adecuada cuando se trabaja con datasets que contienen outliers significativos.

En mi opinión, si los datos tienen distribuciones muy diferentes y hay valores atípicos, la estandarización es más recomendable, ya que no depende de valores mínimos o máximos. Si los datos deben estar dentro de un rango específico, la normalización es la mejor alternativa, especialmente en modelos basados en distancia o en redes neuronales.

**¿Cuál usas más en tus proyectos? ¿Has notado diferencias al aplicar una u otra?**
**Te leo en los comentarios!!** 👨‍💻

**Fuente**:
https://www.linkedin.com/posts/carlosramirezmartin_machinelearning-datascience-featurescaling-activity-7309982353756504064-qt4n?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAjRDYQBAq-lYi8hftkdXAovzZvm7bwZFoA

Carlos Ramírez Martín