12/02/2024

Análisis de la normalidad y normalización de datos

Hay muchas características que necesitan que la distribución de los datos sean una “distribución normal”. Es importante comprobar que las características numéricas siguen una distribución normal.

Hay funciones que te dicen si los datos siguen una distribución normal

Importante: que los datos no tengan una distribución normal no significa que la población no siga una distribución normal.

En ocasiones la suma de distribuciones es lo que da una distribución normal.

Normalización de datos

Es importante que los datos estén entre 0 y 1

Hay diferencia entre normalización y estandarización

* Normalización: pasar de 0 a 1. Python lo hace con MinMaxScaler
* Estandarización: transformar la distribución en un rango de datos cuya media es 0 y su desviación estándar es 1

Siempre que se normaliza se pierde información. No siempre se mejora el modelo al normalizar, pero generalmente merece la pena