[About Feature Scaling and Normalization (sebastianraschka.com)]
[How and why to Standardize your data: A python tutorial | Towards Data Science]
desvio padrão = variância^½
desvio = diferença em módulo de cada termo com a média
variância = média dos quadrados dos desvios

Normalização
O objetivo da normalização é transformar características para estar em uma escala semelhante.

Standard Scaler

Scaling to a range
Scaling to a range significa converter valores de características de ponto flutuante de sua faixa natural (por exemplo, 100 a 900) em uma faixa padrão (por exemplo, 0 a 1 ou -1 a +1). Assim ajudará o modelo a compreender melhor o peso de cada recurso, não dando “valor maior” para aquele recurso que estiver “mais distante” de outros. (idade 2 anos não tem menos valor que a idade 80, porem nao seria ideal utilizar em dados que representam a renda)

$$x' = (x - x_{min})/(x_{max} - x_{min})$$

O Scaling to a range é uma boa escolha quando ambas as seguintes condições são atendidas:

- Você conhece os limites superiores e inferiores aproximados em seus dados com poucos ou nenhum outliers.
- Seus dados são aproximadamente distribuídos uniformemente através desse intervalo.

Standard Scaler
Padronizar os recursos removendo a média e escalonando para a variância da unidade.
O standard score de uma amostra X é calculado como:
z=(x-u)/s
onde **u** é a média das amostras de treinamento e **s** é o desvio padrão.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i) \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- StandardScaler remove a média e dimensiona cada característica/variável para variância unitária. Esta operação é realizada em termos de características de forma independente.

- StandardScaler pode ser influenciado por outliers (se existirem no conjunto de dados) uma vez que envolve a estimativa da média empírica e desvio padrão de cada recurso.

Numpy.log

Log Scaling
A escala de log calcula o log de seus valores para comprimir uma ampla faixa em uma faixa estreita.

$$x' = log(x)$$

É indicado usar quando alguns dados tem muito valores enquanto a maioria não.
Numpy.log
O numpy.log é uma função matemática que calcula o logaritmo natural de um elemento da matriz de entrada.

Clipping
Se o conjunto de dados contiver outliers extremos, você pode tentar o recorte de recursos, que limita todos os valores de recurso acima (ou abaixo) de um determinado valor ao valor fixo.

Z-score
Z-score é uma variação de scaling que representa o número de desvios padrão longe da média. Você usaria z-score para garantir que suas distribuições de recursos tenham mean = 0 e std = 1. É útil quando há alguns outliers, mas não tão extremo que você precisa de clipping.

A fórmula para calcular a pontuação z de um ponto, x,é a seguinte:

$$x' = (x - \mu)/\sigma$$ **->** μ é a média e σ é o desvio padrão.

Técnica	**Fórmula**	**Quando Usar**
Linear Scaling	x’ = (x-x_min)/(x_max-x_min)	Quando os dados são mais ou menos uniformemente distribuído em uma faixa fixa
Clipping	se x>max, então x’=max. se x<min, então x’=min	Quando os dados contém alguns outliers extremos
Log Scaling	x’=log(x)	Quando os dados estão de acordo com a “lei da energia” (uns muitos maioria poucos)
Z-score	x’=(x-u)/s	Quando a distribuição dos dados não contém outliers extremos