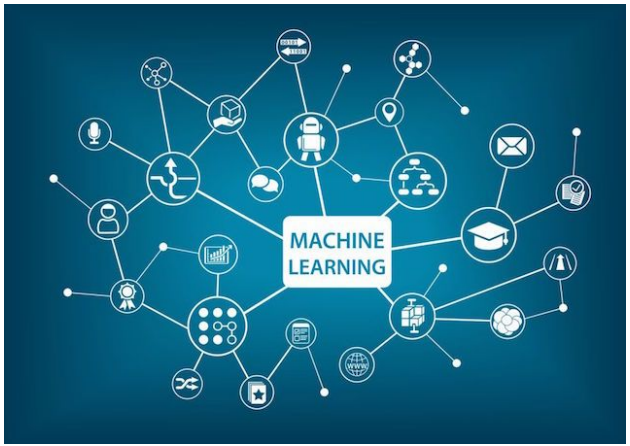
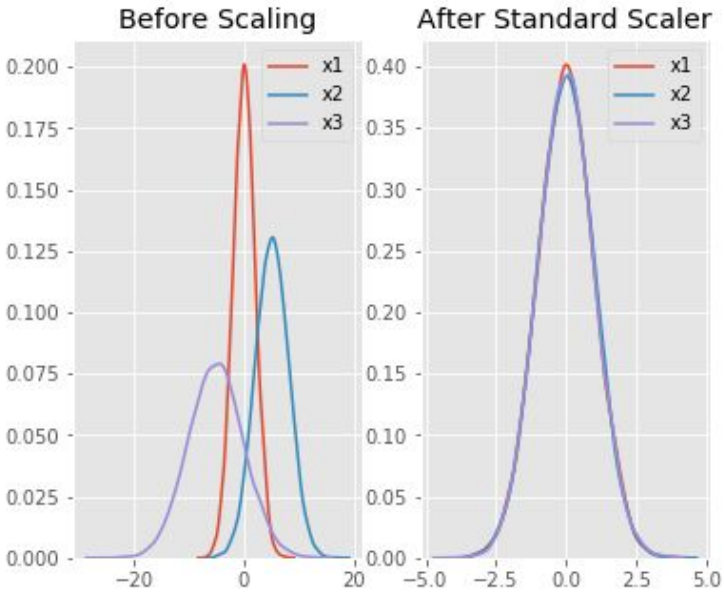

Cientista de Dados
na Prática

O que é ... Por quê usar?

Normalização ou Padronização dos Dados

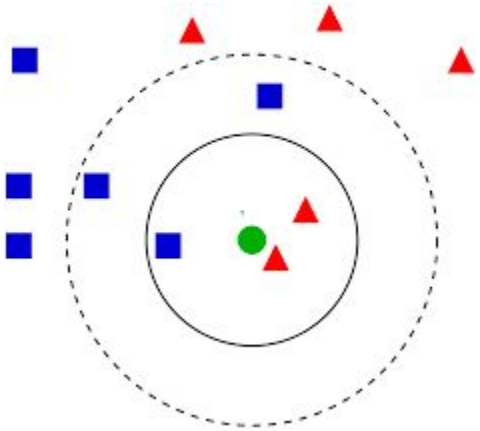


Técnicas para Redimensionar os Dados



Por que devemos usar o dimensionamento de recursos?

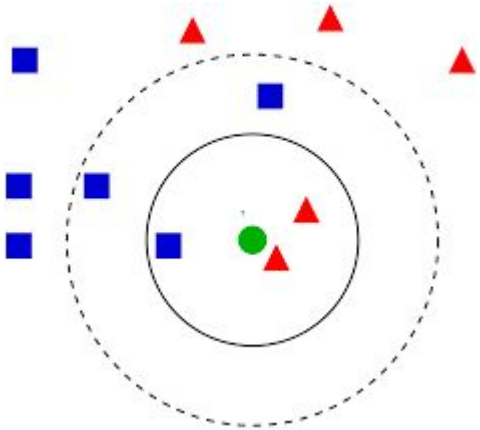
A primeira questão que precisamos abordar - por que precisamos dimensionar as variáveis em nosso conjunto de dados? Alguns algoritmos de aprendizado de máquina são sensíveis ao dimensionamento de recursos, enquanto outros são virtualmente invariantes a ele. Deixe-me explicar isso com mais detalhes.



Algoritmos baseados em distância

Algoritmos de distância como KNN , K-means e SVM são os mais afetados pela variedade de recursos. Isso ocorre porque, nos bastidores, **eles usam distâncias entre pontos de dados para determinar sua similaridade.**

Por exemplo, digamos que temos dados contendo pontuações de alunos no CGPA do ensino médio (variando de 0 a 5) e suas receitas futuras (em milhares de rúpias):



	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

Uma vez que os dois recursos têm escalas diferentes, há uma chance de que uma ponderação mais alta seja dada aos recursos de maior magnitude. Isso afetará o desempenho do algoritmo de aprendizado de máquina e, obviamente, não queremos que nosso algoritmo seja inclinado para um recurso.

> *Portanto, dimensionamos nossos dados antes de empregar um algoritmo baseado em distância, de modo que todos os recursos contribuam igualmente para o resultado.*

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

O efeito da escala é notável quando comparamos a distância euclidiana entre os pontos de dados para os alunos A e B, e entre B e C, antes e depois da escala, conforme mostrado abaixo:

- Distância AB antes de escalar => $\sqrt{(40-60)^2+(3-3)^2}=20$
- Distância BC antes de escalar => $\sqrt{(40-40)^2+(4-3)^2}=1$
- Distância AB após escalar => $\sqrt{(1.1+1.5)^2+(1.18-1.18)^2}=2.6$
- Distância BC após escalar => $\sqrt{(1.1-1.1)^2+(0.41+1.18)^2}=1.59$

O dimensionamento trouxe os recursos para a imagem e as distâncias agora são mais comparáveis do que eram antes de aplicarmos o dimensionamento.

O que é normalização?

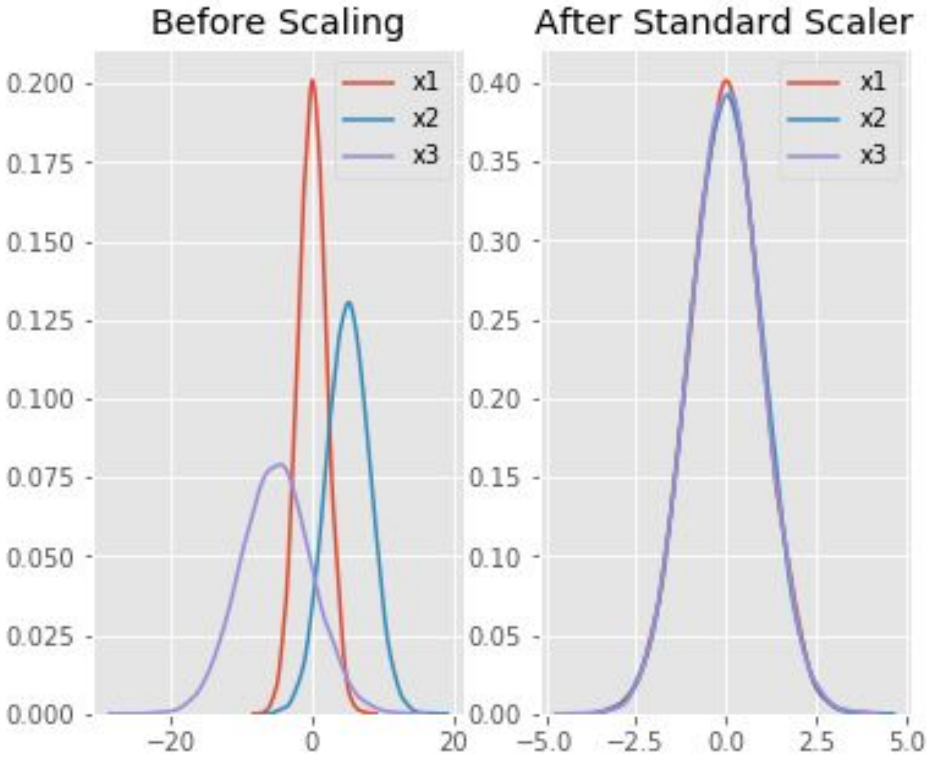
A normalização é uma técnica de escala na qual os valores são deslocados e redimensionados para que fiquem entre 0 e 1. Também é conhecida como escala Mín-Máx.

Aqui está a fórmula para normalização:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Aqui, Xmax e Xmin são os valores máximo e mínimo do recurso, respectivamente.

- Quando o valor de X é o valor mínimo na coluna, o numerador será 0 e, portanto, X 'é 0
- Por outro lado, quando o valor de X é o valor máximo da coluna, o numerador é igual ao denominador e, portanto, o valor de X 'é 1
- Se o valor de X estiver entre o valor mínimo e máximo, então o valor de X 'está entre 0 e 1


Before Scaling
After Standard Scaler
x1
x2
x3
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−20
0
20
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−5.0
−2.5
0.0
2.5
5.0

O que é padronização?

A padronização é outra técnica de dimensionamento em que os valores são centralizados em torno da média com um desvio padrão da unidade. Isso significa que a média do atributo torna-se zero e a distribuição resultante tem um desvio padrão da unidade.

Esta é a fórmula para padronização:

$$X' = \frac{X - \mu}{\sigma}$$

μ é a média dos valores do recurso e σ é o desvio padrão dos valores do recurso. Observe que, neste caso, os valores não estão restritos a um intervalo específico.

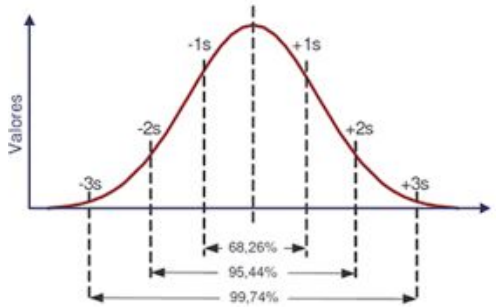
Agora, a grande questão em sua mente deve ser quando devemos usar a normalização e quando devemos usar a padronização? Vamos descobrir!

Exemplo Prático por Gentileza....

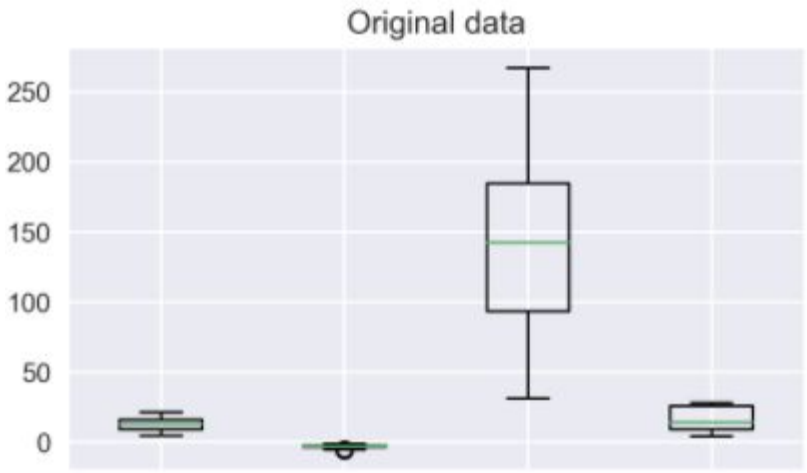
A grande questão - normalizar ou padronizar?

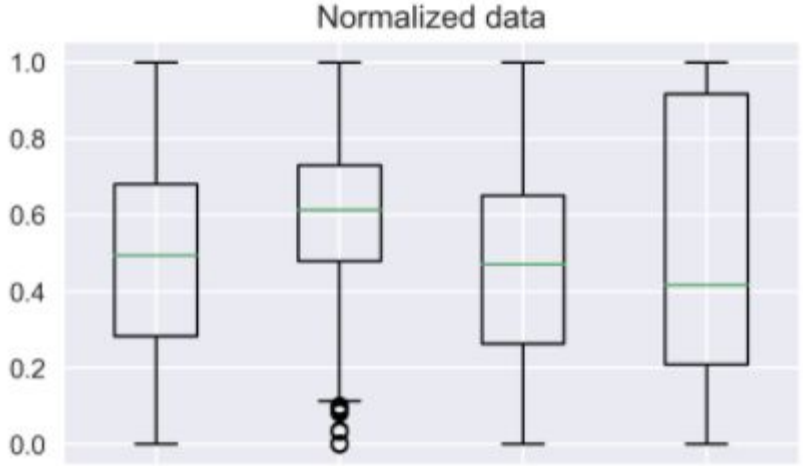
alguns escolhem....

- A normalização é boa para usar quando você sabe que a distribuição de seus dados não segue uma distribuição Gaussiana. Isso pode ser útil em algoritmos que não assumem nenhuma distribuição de dados, como K-vizinhos mais próximos e redes neurais.
- A padronização, por outro lado, pode ser útil nos casos em que os dados seguem uma distribuição gaussiana. No entanto, isso não precisa ser necessariamente verdade. Além disso, ao contrário da normalização, a padronização não tem um intervalo delimitador. Portanto, mesmo que você tenha valores discrepantes em seus dados, eles não serão afetados pela padronização.

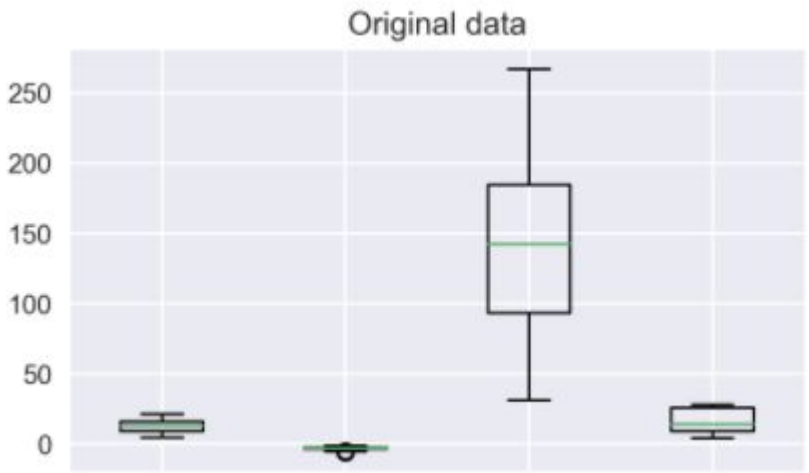


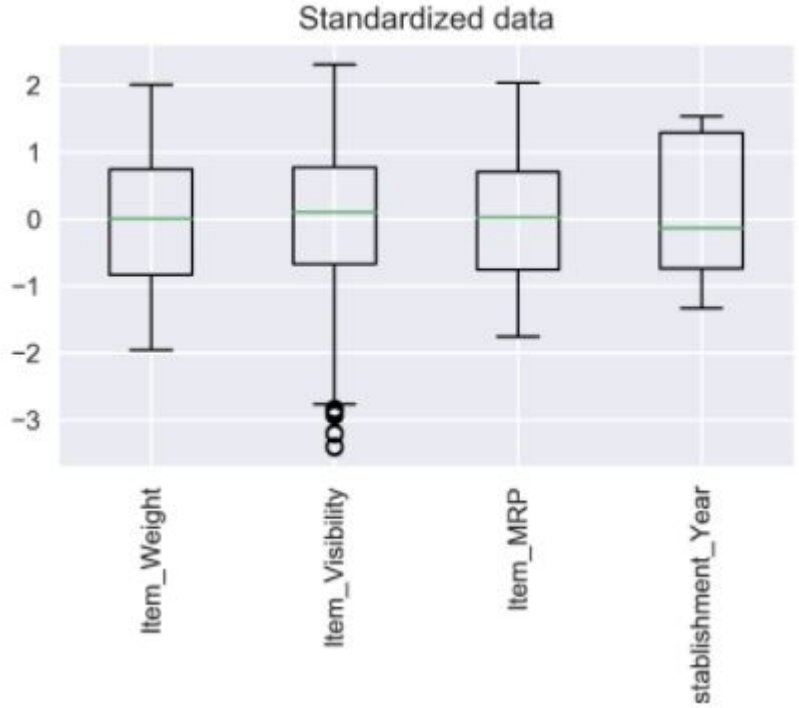
Passam a ser comparáveis





Passam a ser comparáveis





Qual escolher?

	RMSE
Original	1319.283626
Normalized	1174.205859
Standardized	1183.448734

O que der o melhor resultado

Lembre-se de que **não há uma resposta correta** para quando usar normalização em vez de padronização e vice-versa.

Tudo depende dos seus dados e do algoritmo que você está usando.

Testar ....

Simbóra

....