Categorical Encoding

- Algoritmos entendem números
- *Categorical encoding é o processo de transformar categorias em números*
- *Duas Formas:*
  - Label encoding
  - One-hot encoding

Label encoding

- Cada categoria recebe um número, normalmente em ordem alfabética

Label encoding

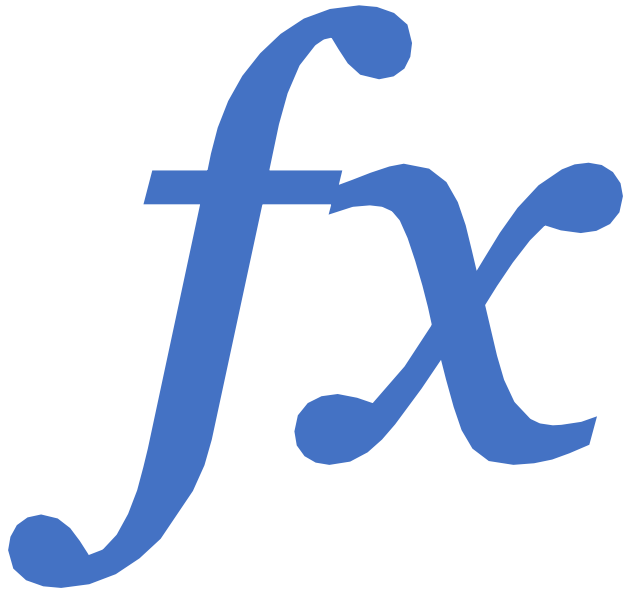
TIPOSBENS
bens imóveis
bens imóveis
bens imóveis
seguro de vida
nenhuma
nenhuma
seguro de vida
carro
bens imóveis
carro
carro
seguro de vida
carro
carro
carro
carro
seguro de vida
carro
nenhuma
carro
carro

bens imóveis	0
seguro de vida	1
nenhuma	2
carro	3

TIPOSBENS
0
0
0
1
2
2
1
3
0
3
3
1
3
3
3
3
1
3
2
3
3

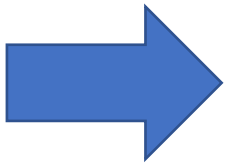
One-hot encoding

- Cada categoria é transformada em outro atributo: dummy variable
- Um valor binário informa a ocorrência



One-hot encoding

TIPOSBENS
bens imóveis
bens imóveis
bens imóveis
seguro de vida
nenhuma
nenhuma
seguro de vida
carro
bens imóveis
carro
carro
seguro de vida
carro
carro
carro
carro
seguro de vida
carro
nenhuma
carro
carro



bens imóveis	seguro de vida	nenhuma	carro
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
0	0	1	0
0	0	1	0
1	0	0	0
0	0	0	1
1	0	0	0
0	0	0	1
0	0	0	1
0	1	0	0
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1
0	1	0	0
0	0	0	1
0	0	1	0
0	0	0	1
0	0	0	1

Qual valor?

bens imóveis	seguro de vida	nenhuma	carro
1	?	?	?
1	?	?	?
1	?	?	?
1	?	?	?

Dummy Variable Trap

- O valor dos atributos se torna altamente previsível
- Resultado, correlação entre as variáveis Independentes: multicolinearidade
- Solução: Excluir um dos atributos!

bens imóveis	seguro de vida	nenhuma	carro
1	?	?	?
	?	?	?
1	?	?	?
1	?	?	?

Qual usar?

Label encoding	One-hot encoding
Há ordem (progr. Junior, Pleno, Sênior)	Não há ordem
Grande Número de categorias, não da pra usar One-hot encoding	Número de categorias é pequeno

Dimensionamento de Características

- Processo de transformação de dados numéricos
- Variáveis em escalas diferentes
  - Contribuem de forma desbalanceada para o modelo
  - Exemplo: Salário e Altura
- Gradient Descent converge mais rapidamente para o mínimo local

Dimensionamento de Características

- Padronização (Z-score)
- Normalização (Min-Max)

Padronização (Z-score)

- Dados aproximados da média (zero) e desvio padrão 1
- Podem ser negativos
- Não afeta outliers
- Deve ser usado na maioria dos casos

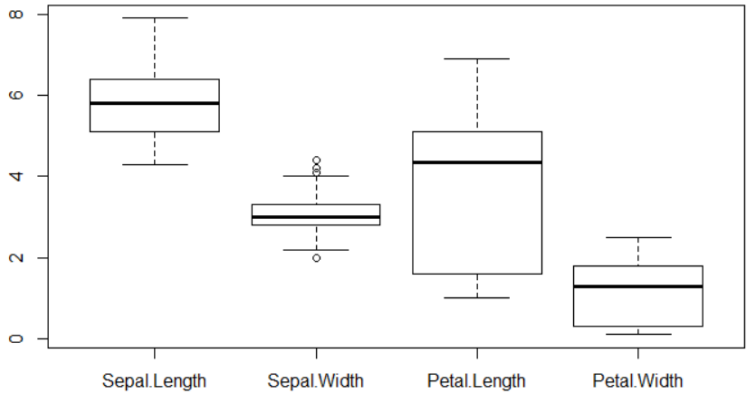
Normalização (Min-Max)

- Transforma para escala comum entre zero e 1
- Usado em processamento de imagens e RNA
- Quando não sabemos a distribuição dos dados
- Quando precisam ser positivos
- Algoritmos não "requerem" dados normais
- Remove outliers pois impõe "limites"

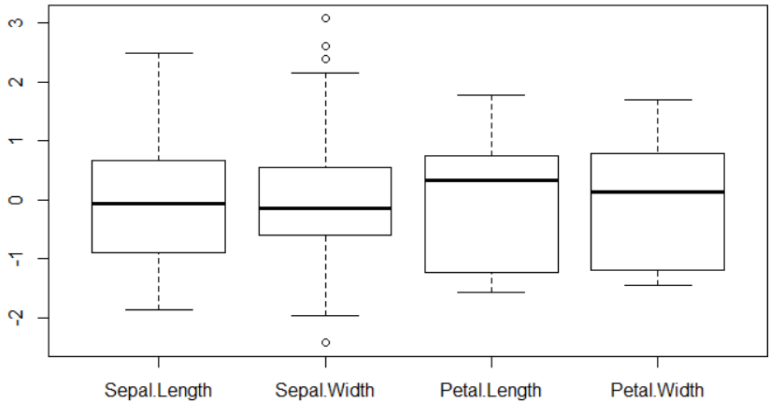
- $X_n = \frac{X - X_{min}}{X_{max} - X_{min}}$

IRIS

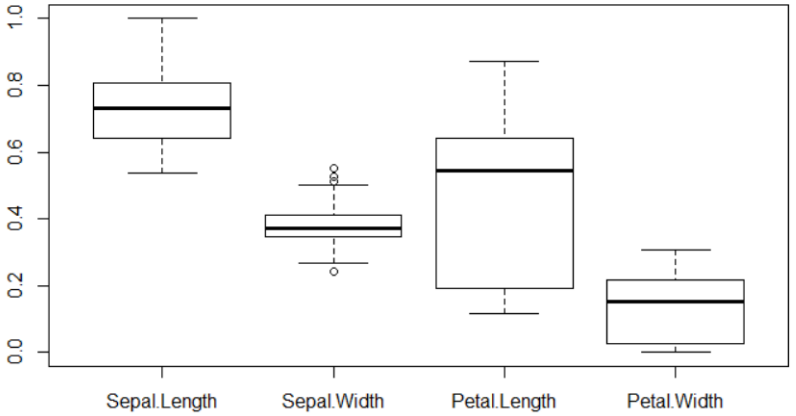
IRIS



Padronização (Z-score)



Normalização (min-max)



Dimensionamento de Características

- Não vai necessariamente melhorar seu modelo!
- Arvores de decisão não precisam de nenhum tipo
- Não se aplica a atributos categóricos transformados