

Formação Cientista de Dados

Codificação de Categorias



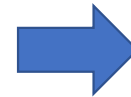
Categorical Encoding

- 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

EstadoCivil
Casado
Solteiro
Divorciado
Casado
Solteiro
Casado
Solteiro
Casado
Casado
Solteiro
Casado
Solteiro
Divorciado
Casado
Solteiro
Casado
Casado
Casado
Solteiro



Casado	0
Divorciado	1
Solteiro	2



EstadoCivil
0
1
2
0
2
0
2
0
0
2
0
2
2
1
0
2
0
0
0
0
2

Label encoding


- Problema: o algoritmo pode correlacionar os dados como uma ordem de grandeza!

Fidelidade
Silver
Gold
Silver
Silver
Silver
Silver
Gold
Silver
Silver
Silver
Silver
Silver
Silver
Gold
Silver
Silver
Silver
Platinum
Silver

EstadoCivil
Casado
Solteiro
Divorciado
Casado
Solteiro
Casado
Solteiro
Casado
Casado
Solteiro
Casado
Solteiro
Divorciado
Casado
Solteiro
Casado
Casado
Casado
Solteiro

One-hot encoding

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



EstadoCivil
Casado
Solteiro
Divorciado
Casado
Solteiro
Casado
Solteiro
Casado
Solteiro
Casado
Solteiro
Casado
Solteiro
Divorciado
Casado
Solteiro
Casado
Casado
Casado
Solteiro

Casado	Solteiro	Divorciado
1	0	0
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
0	1	0
1	0	0
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
1	0	0
1	0	0
0	1	0

Qual valor?

Casado	Solteiro	Divorciado
1	?	?
0	?	0
0	0	?
1	?	?



Dummy Variable Trap

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

Casado	Solteiro	Divorciado
1	0	0
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
0	1	0
1	0	0
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
1	0	0
1	0	0
0	1	0





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