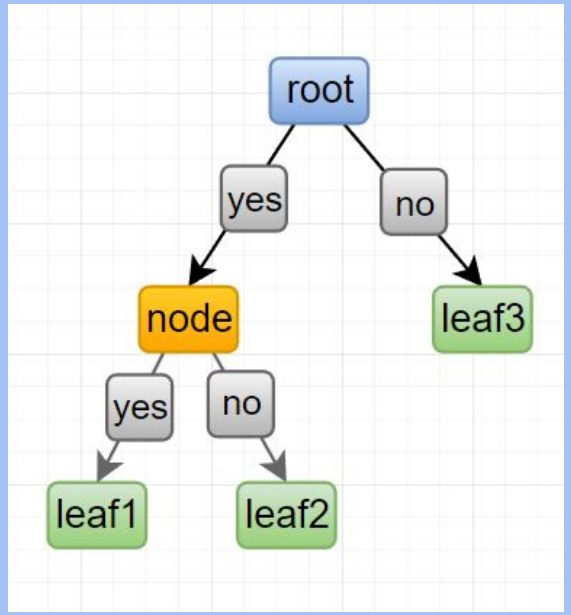
Introdução

Árvores de decisão

Temas principais:
Classificação em árvore de decisão
Regressão em árvore de decisão
Floresta aleatória (classificação e regressão)



Revisão

Tipo de variável

Variáveis Quantitativas: valores numéricos

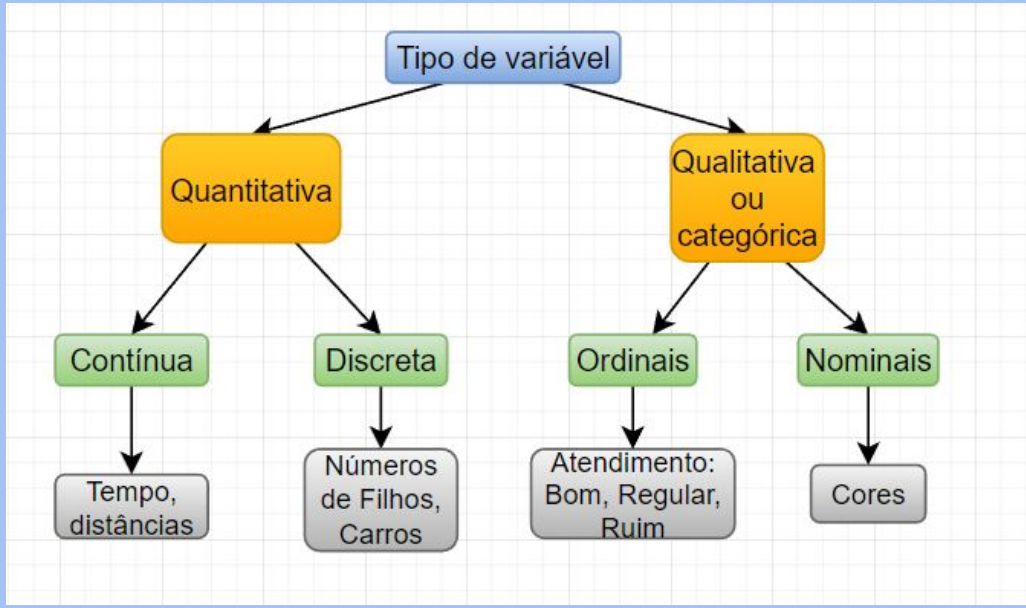
Variáveis Contínuas: escala contínua

Variáveis Discreta: valores inteiros

Variáveis Qualitativas (ou categóricas): várias categorias

Variáveis Ordinais: existe uma ordenação

Variáveis Nominais: não existe ordenação



http://leg.ufpr.br/~silvia/CE055/node8.html#:~:text=Vari%C3%A1veis%20cont%C3%ADnuas%2C%20caracter%C3%ADsticas%20mensur%C3%A1veis%20que,)%2C%20press%C3%A3o%20arterial%2C%20idade.

Qual a utilidade?

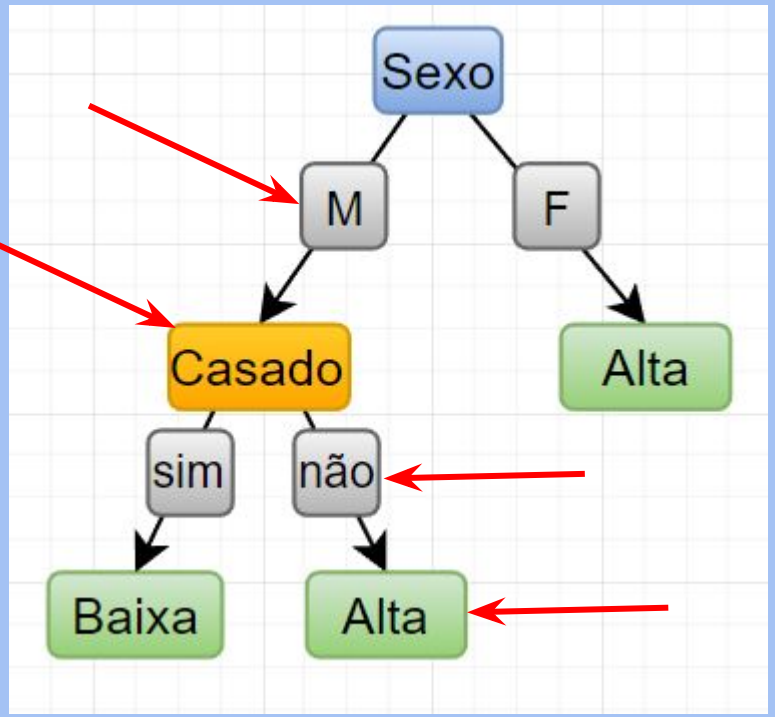
Classificação:
Estimar e prever categorias
(Vivo ou morto?)

Regressão:
Estimar e prever valores
(Qual sua idade?)

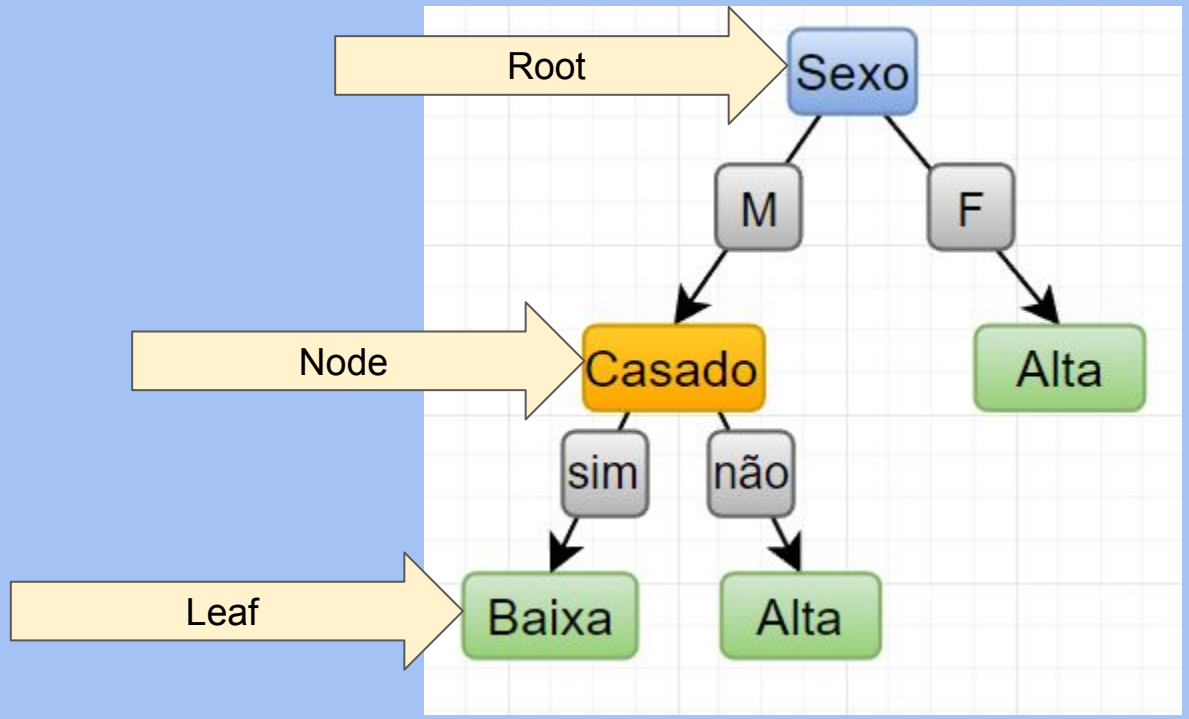
Prevendo

Qual a chance, de um homem solteiro, comprar o nosso perfume feminino?

Alta



Estrutura



Árvore de decisão(Classificação)

Como criar uma árvore de decisão(Classificação)?

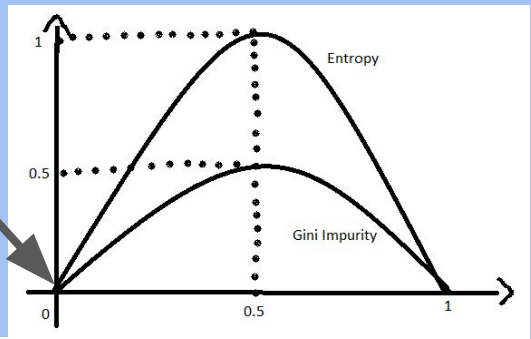
1° Selecionando colunas usando o gini impurity ou entropia:

Dados		
Sapato	**Óculos**	**Chance**
sim	sim	Alta
não	sim	Baixa
sim	não	Baixa
não	não	Baixa

Escolher a coluna sapatos?

Escolher o menos impuro!

Ou escolher a coluna óculos?



$$Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$$

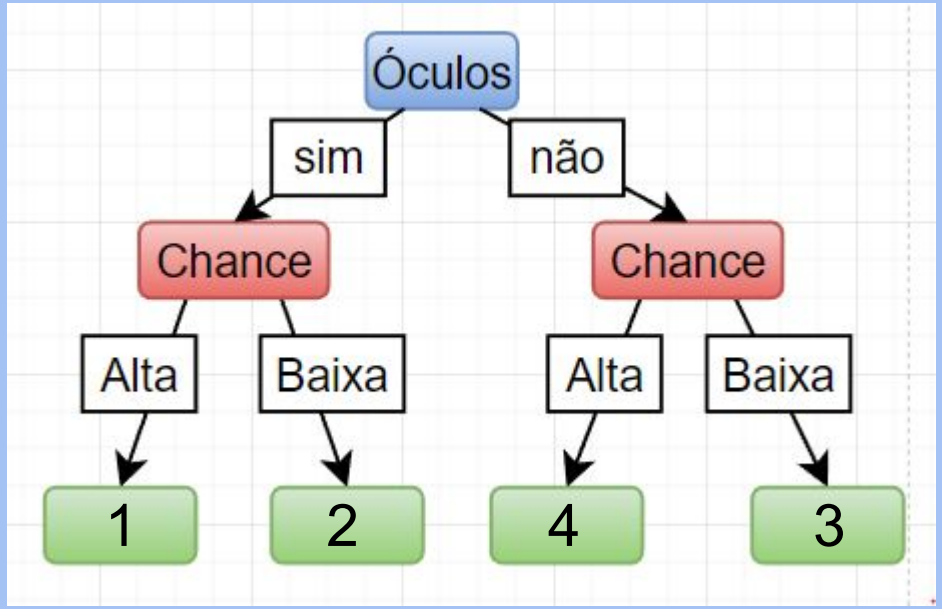
$$Entropy = \sum_{i=1}^{n} -p(c_i)log_2(p(c_i))$$

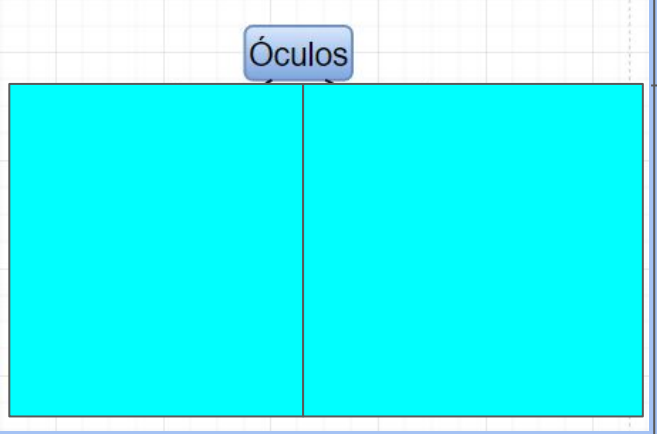
where $p(c_i)$ is the probability/percentage of class c_i in a node.

https://www.geeksforgeeks.org/gini-impurity-and-entropy-in-decision-tree-ml/
https://medium.com/thatascience/gini-index-vs-entropy-for-information-gain-in-decision-trees-252f9afa8229

Calculando gini impurity

Binário	
Óculos	Chance
sim	Alta
não	Baixa
sim	Baixa
não	Alta
não	Alta
não	Alta
não	Baixa
não	Alta
não	Baixa
sim	Baixa





Total Gini = P1.W1 + P2.W2

P1= Probabilidade do “sim” = (1+2)/10 = 0,3

W1= Gini Impurity do “sim” = 0,4444

P2= Probabilidade do “não” = (4+3)/10 = 0,7

W2= Gini Impurity do “não” = 0,4898

Total Gini =0,4444.0,3 + 0,4898.0,7

Total Gini = 0,47618

Calculando gini impurity

do sim

Gini Impurity = 1 - (Probabilidade do “Alta”)² - (Probabilidade do “Baixa”)²

Probabilidade do “Alta”= $\frac{1}{1+2}$ Probabilidade do “Baixa”= $\frac{2}{1+2}$

Gini Impurity = 1 - (1/(1+2))² - (2/(1+2))²

Gini Impurity = 0,4444

do não

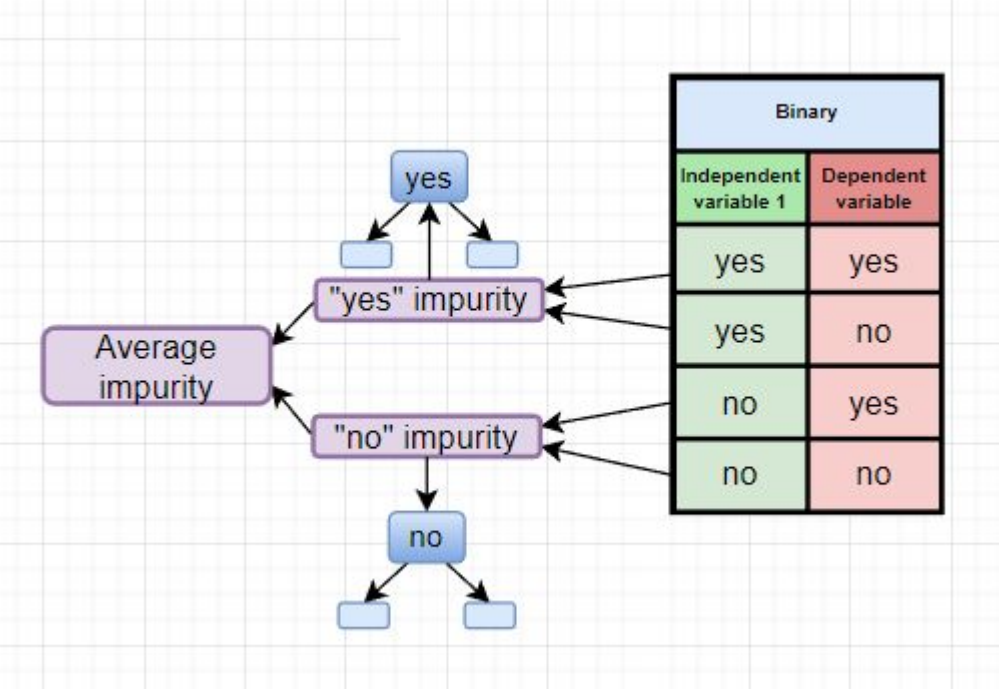
Gini Impurity = 1 - (Probabilidade do “Alta”)² - (Probabilidade do “Baixa”)²

Probabilidade do “Alta”= $\frac{4}{4+3}$ Probabilidade do “Baixa”= $\frac{3}{4+3}$

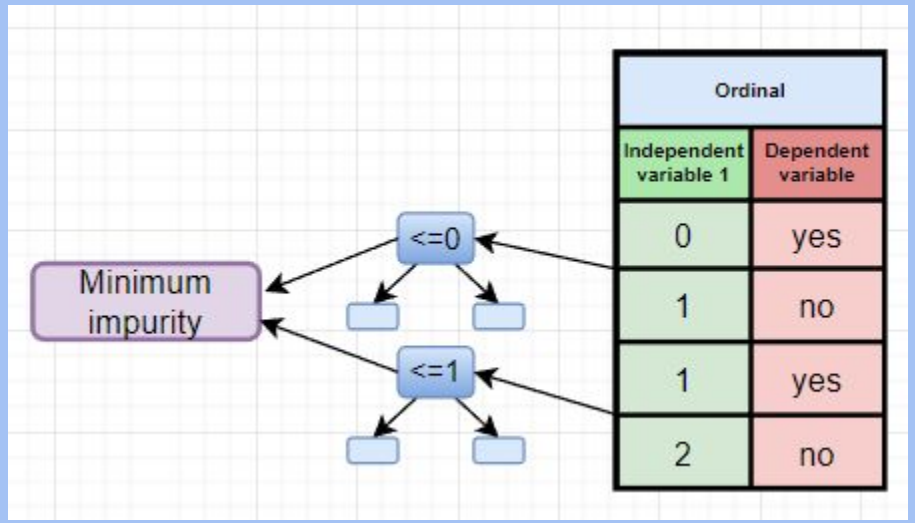
Gini Impurity = 1 - (4(4+3))² - (3/(4+3))²

Gini Impurity = 0,4898

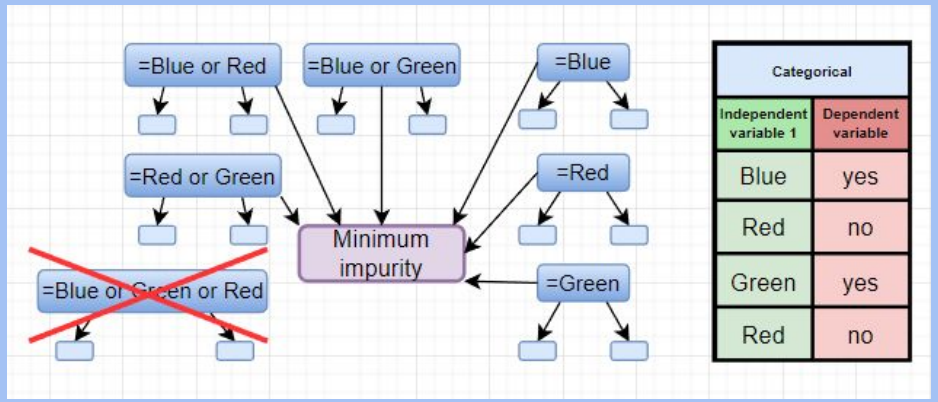
Calcular impureza para diversos tipos de variáveis



Calcular impureza para diversos tipos de variáveis

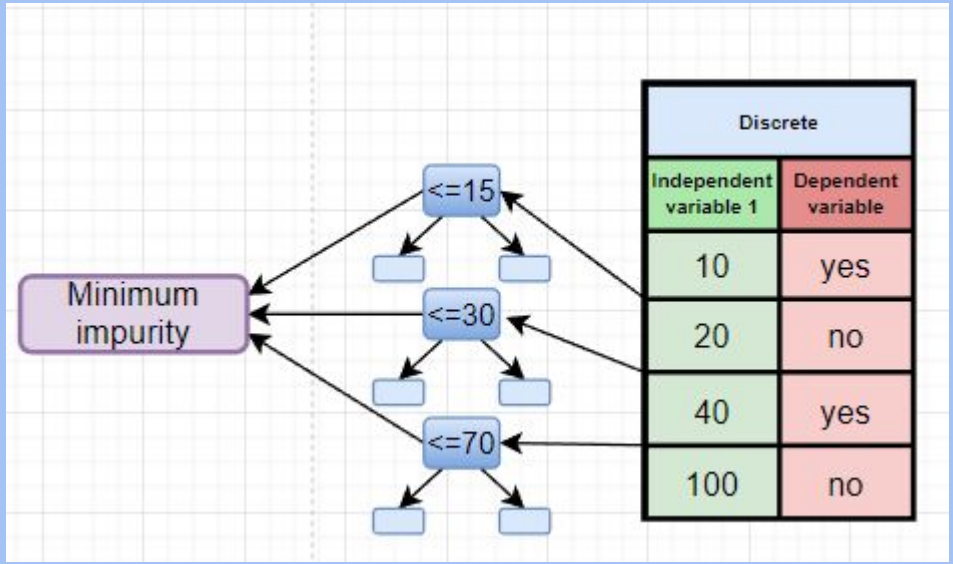


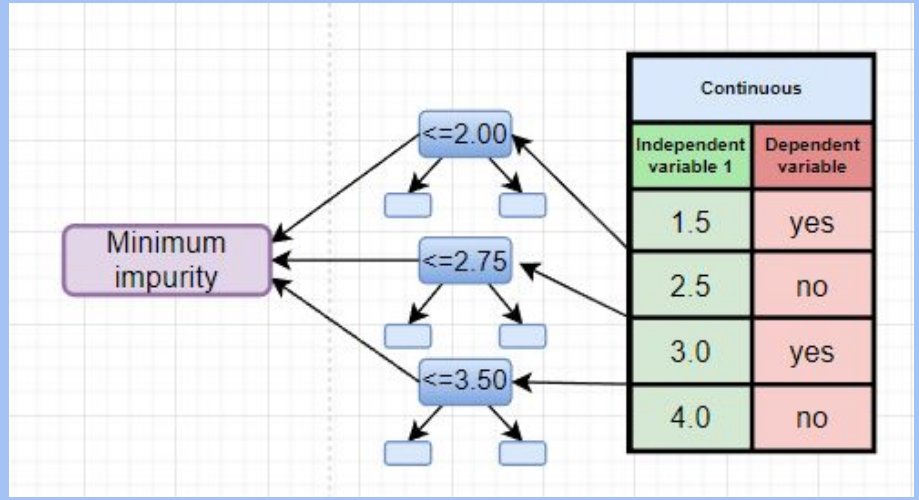
Ordinal	
Independent variable 1	Dependent variable
0	yes
1	no
1	yes
2	no



Categorical	
Independent variable 1	Dependent variable
Blue	yes
Red	no
Green	yes
Red	no

Calcular impureza para diversos tipos de variáveis

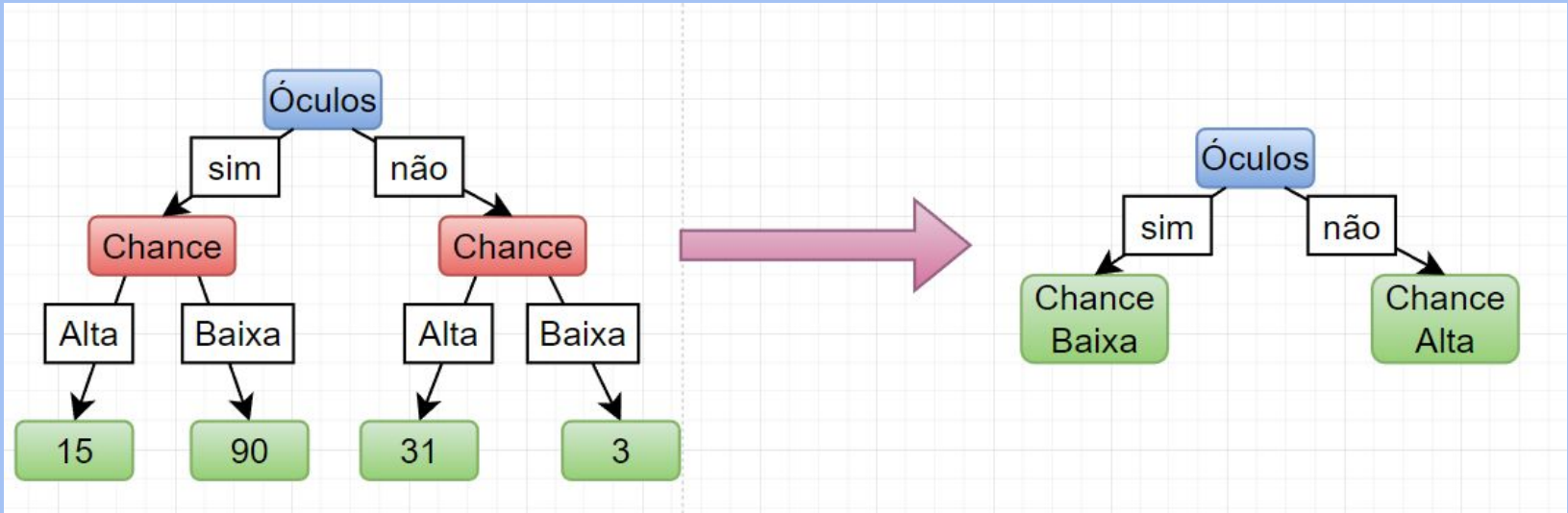




Como criar uma árvore de decisão(Classificação)?

Quando a árvore termina? Ou seja, ao invés de virar um node, vira leaf?

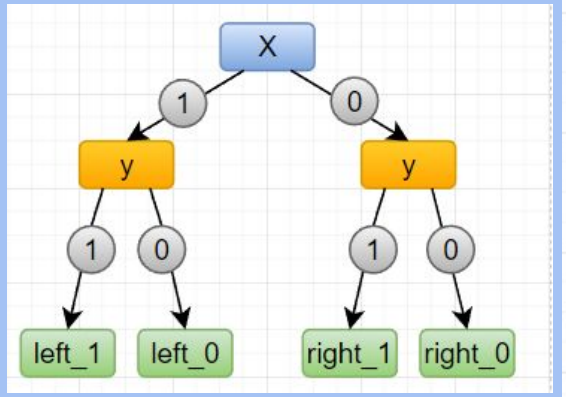
Limites:
-Profundidade máxima da árvore
-Quantidade de amostras que para virar um leaf

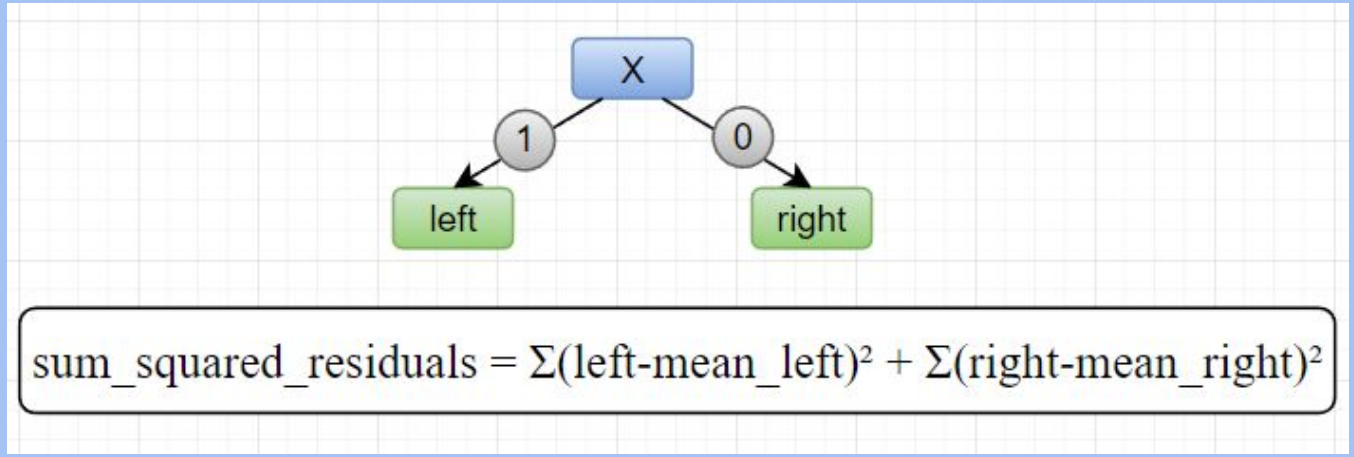


Árvore de decisão(Regressão)

Como criar uma árvore de decisão(Regressão)?

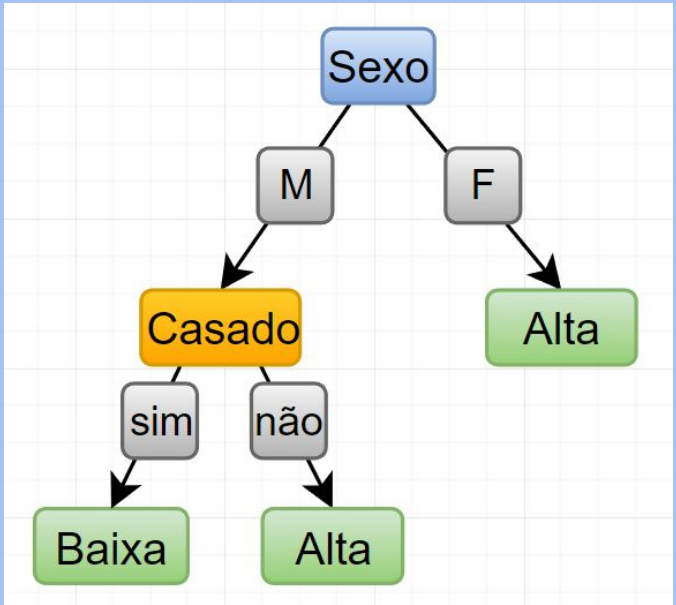
Selecionando colunas	
Classificação	Regressão
Mínima impureza (Gini, entropy)	Mínimo da soma do quadrado dos resíduos

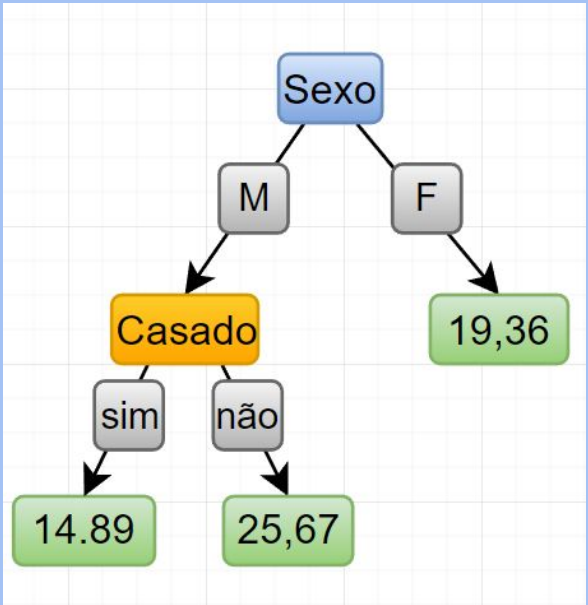




Como criar uma árvore de decisão(Regressão)?

Previsão na Folha (Leaf)	
Classificação	Regressão
Valor categórico (sim, não)	Média dos valores restantes

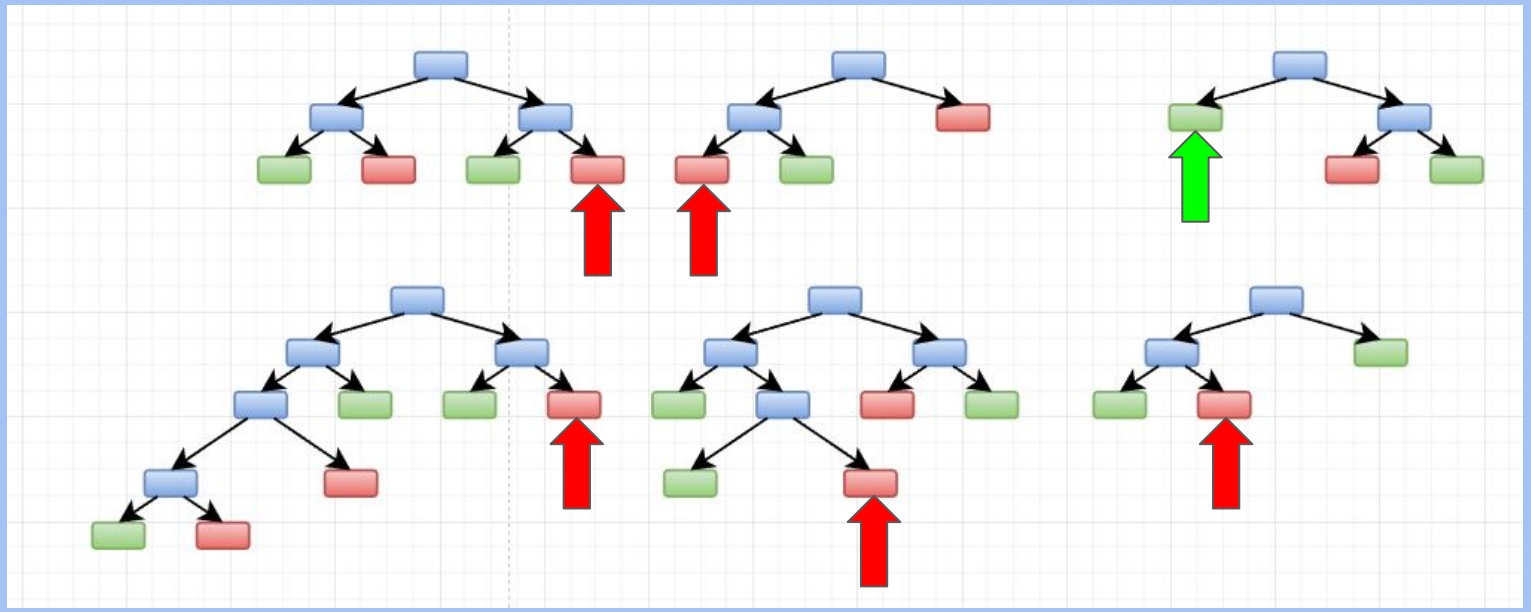




Floresta aleatória

Previsão em uma Floresta aleatória?

Votação	
5 árvores deram vermelho	1 árvore deu verde
RESULTADO: **Vermelho**	

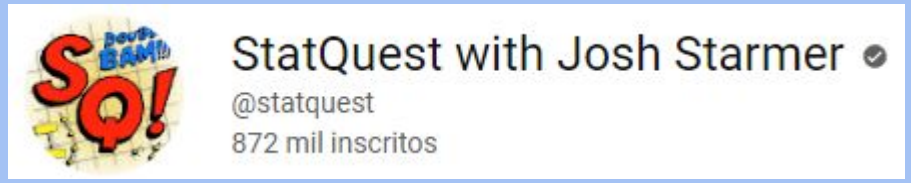


Como criar uma Floresta aleatória?

Seleção de dados		
	Coluna	**Índice (Linha)**
Árvore (Classificação)	gini / entropia	Todas
Árvore (Regressão)	Mínimo da soma do quadrado dos resíduos	Todas
Floresta aleatória	Aleatório em cada node (Podem ser aleatório ou não)	Diferentes amostragens em cada árvore (São sempre aleatórias)

Referências

https://github.com/HygorSantiago/GEIA/tree/main/F%C3%A1bio%20Lofredo/ML-Without-Sklearn



https://youtu.be/_L39rN6gz7Y

https://youtu.be/g9c66TUylZ4

https://youtu.be/J4Wdy0Wc_xQ

https://youtu.be/sQ870aTKqiM

http://leg.ufpr.br/~silvia/CE055/node8.html#:~:text=Vari%C3%A1veis%20cont%C3%ADnuas%2C%20caracter%C3%ADsticas%20mensur%C3%A1veis%20que,)%2C%20press%C3%A3o%20arterial%2C%20idade.

https://www.geeksforgeeks.org/gini-impurity-and-entropy-in-decision-tree-ml/

https://medium.com/thatascience/gini-index-vs-entropy-for-information-gain-in-decision-trees-252f9afa8229