



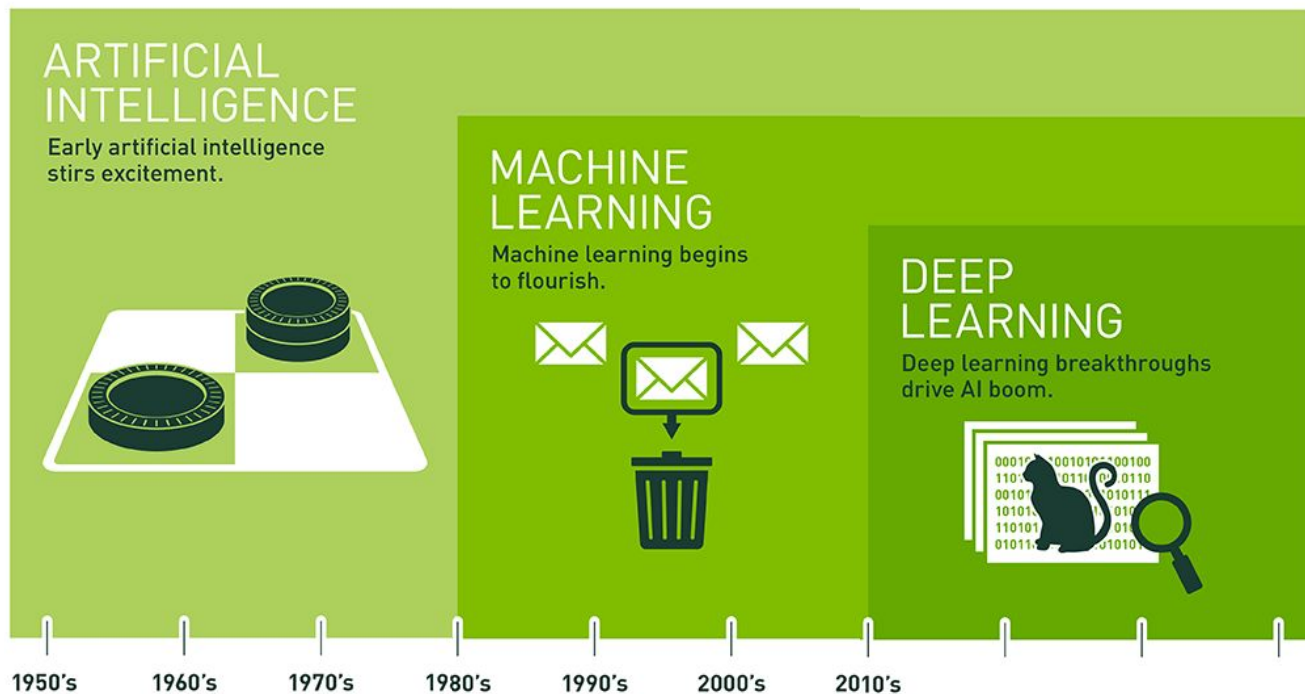
APRENDIZADO DE MÁQUINA - CLASSIFICAÇÃO

profs. Lívia Ruback, Raimundo Macário e Marcelo Dib

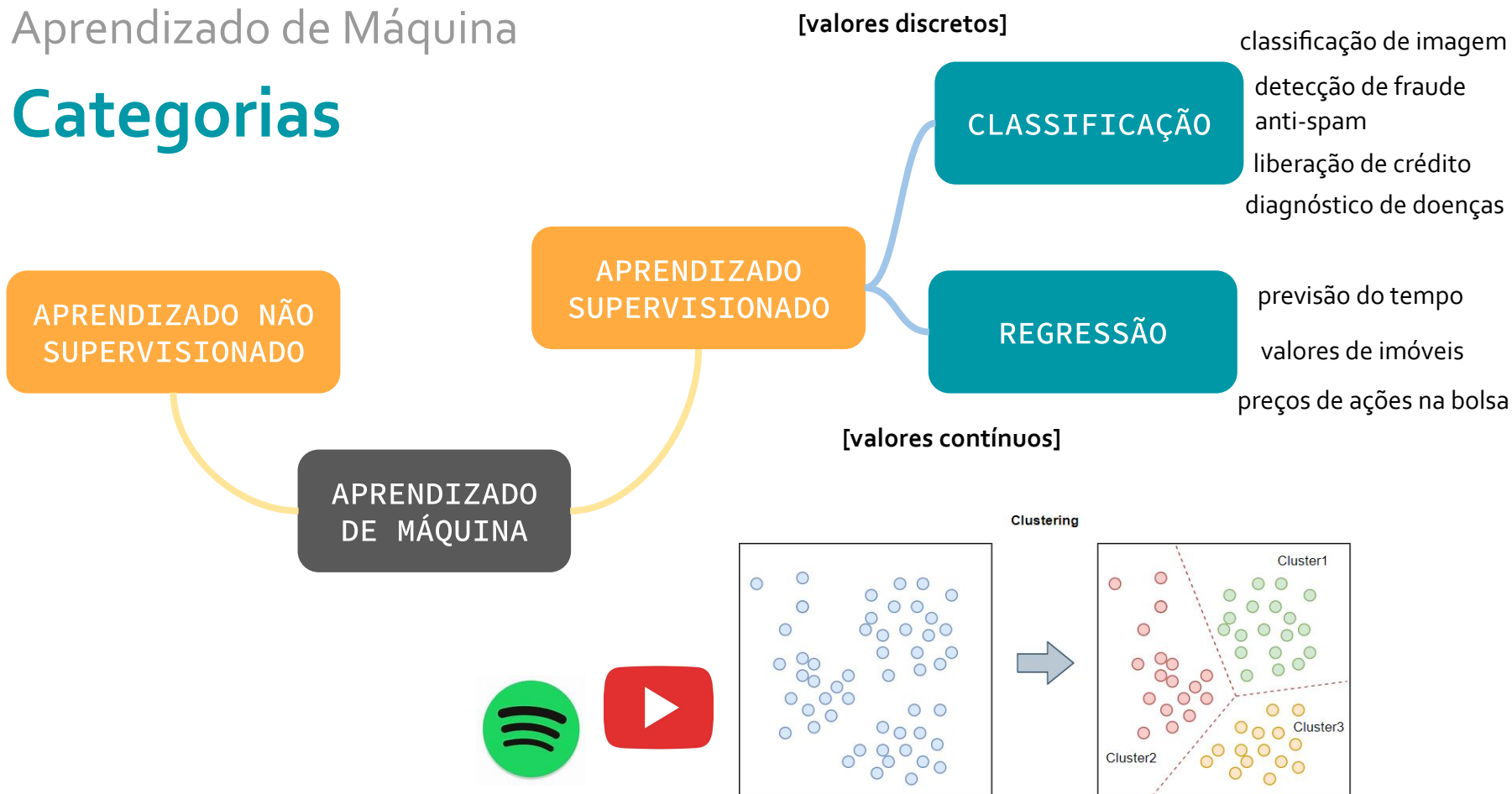
Aprendizado de Máquina - Classificação

- Perspectiva Histórica
- Categorias de Aprendizado de Máquina
- Aprendizado Supervisionado
- Medindo o desempenho
- Valores Categóricos x Valores Contínuos
- Técnicas de Classificação
- Árvores de Decisão
- Entropia
- Parte prática

Perspectiva histórica



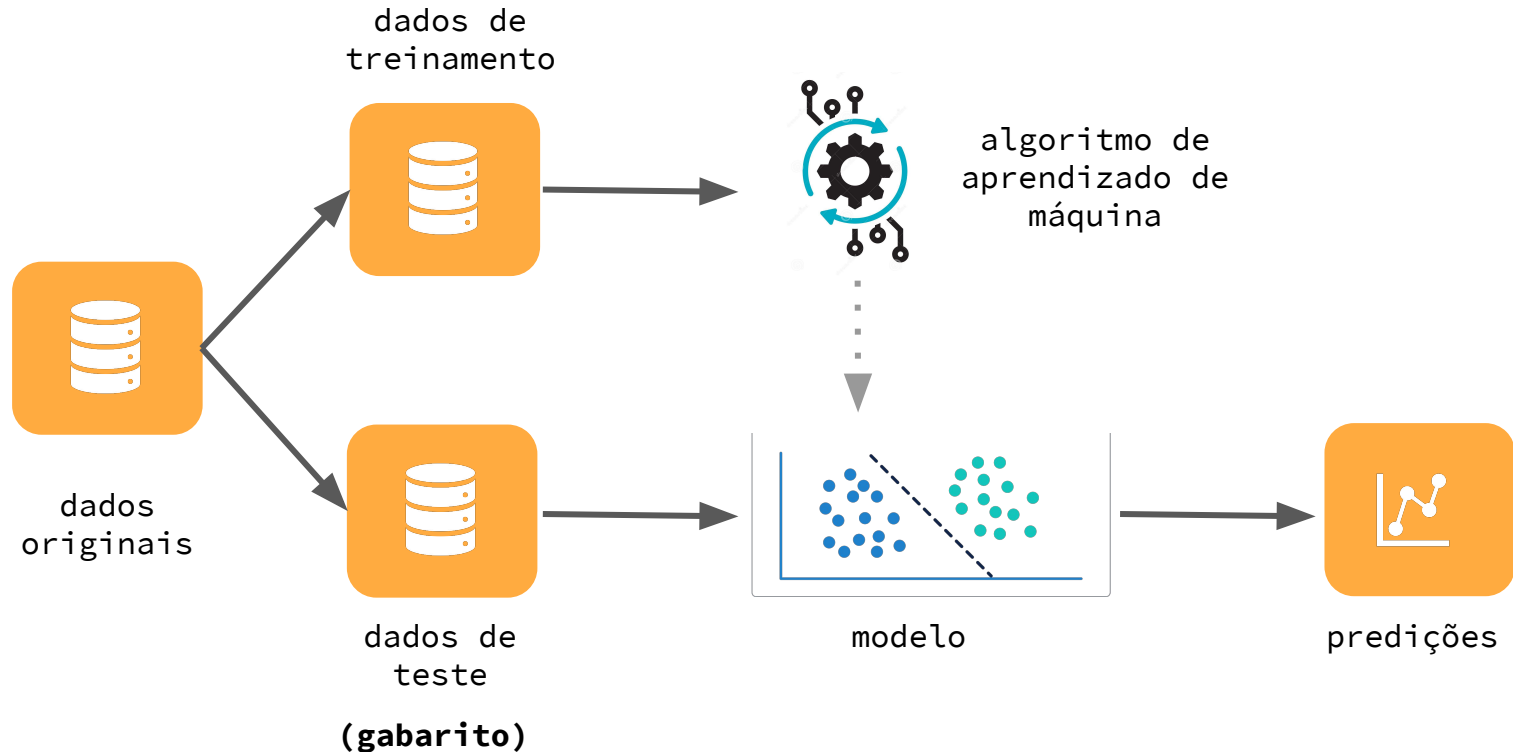
Categorias



Aprendizado Supervisionado

- Quais dados eu tenho?
- Quais o cenário do problema?
- Prevendo valores discretos?
 - Valores categóricos (Sim/Não ou classes)
 - Exemplos: É fraude ou não, Spam ou não, Cachorro ou gato?
- Prevendo valores contínuos?
 - Exemplos: Valor de uma ação, valor do metro quadrado de um imóvel

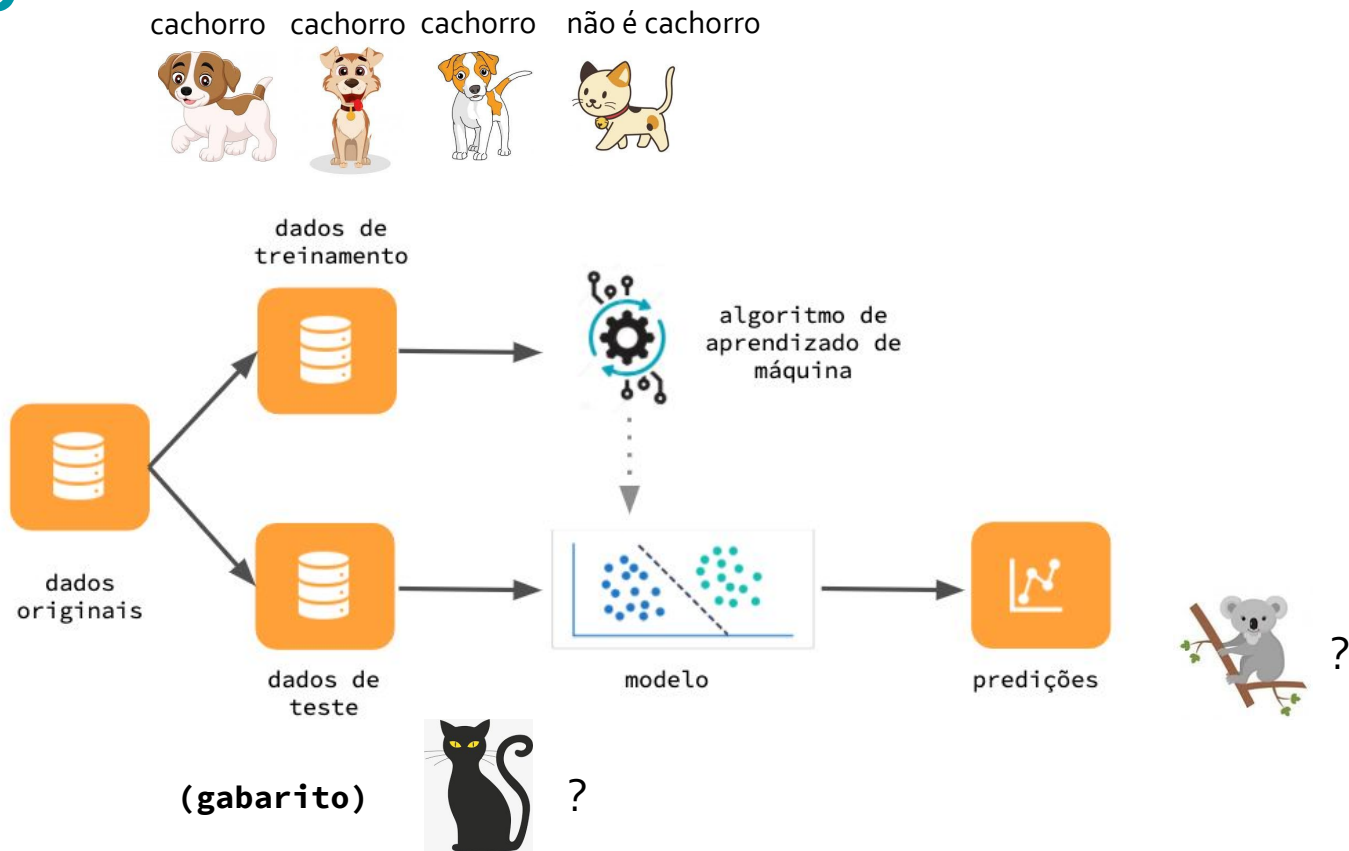
A.M Supervisionado



Aprendizado de Máquina Supervisionado

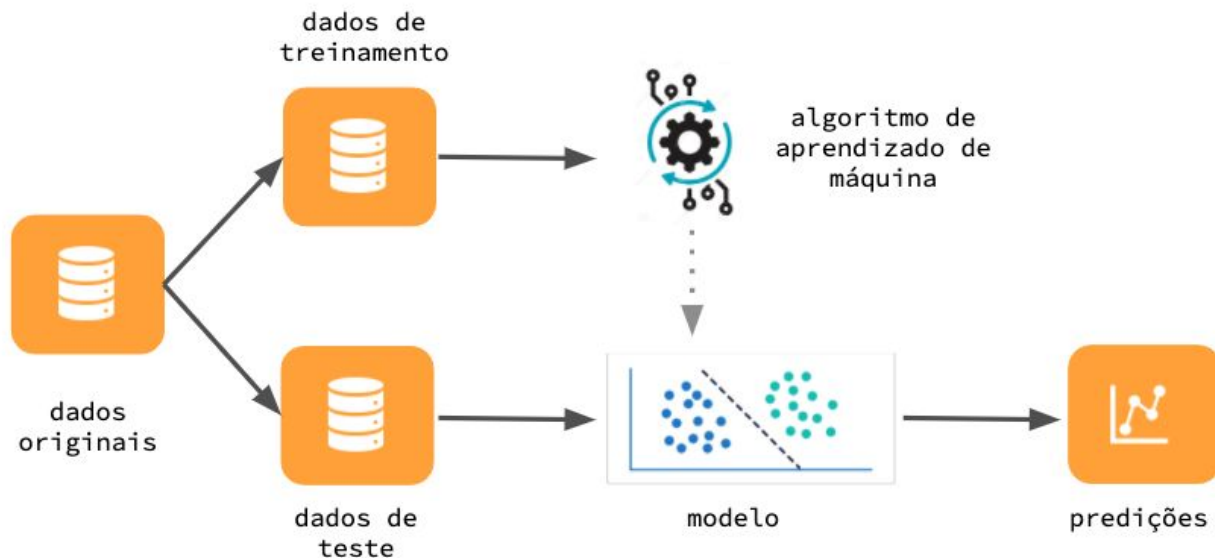
Exemplo

exemplo de classificação!



Medindo o desempenho

como medir o
desempenho do modelo
criado?



Métricas de desempenho

No geral, o quão frequente o classificador está correto?

CLASSIFICAÇÃO DO MODELO

		CLASSIFICAÇÃO DO MODELO			
		cachorro	não é cachorro		
REAL	cachorro	40 VP	10 FN	acertos	erros
	não é cachorro	35 FP	25 VN	VP - Verdadeiros Positivos	VN - Verdadeiros Negativos
				FP - Falsos Positivos	FN - Falsos Negativos

$$\frac{40 \text{ (VP)} + 25 \text{ (VN)}}{40 \text{ (VP)} + 10 \text{ (FN)} + 35 \text{ (FP)} + 25 \text{ (VN)}} = 0.59$$

Acurácia de 59%

Métricas de desempenho

Acurácia

59%

Das imagens que eu classifiquei como cachorros,
quantas de fato continham cachorros?

PREVISTO

cachorro não é
cachorro

REAL

cachorro

não é
cachorro

	cachorro	não é cachorro
cachorro	40 VP	10 FN
não é cachorro	35 FP	25 VN

acertos

erros

VP - Verdadeiros Positivos

VN - Verdadeiros Negativos

FP - Falsos Positivos

FN - Falsos Negativos

40 (VP)

= 0.53

40 (VP) + 35 (FP)

Precisão de 53%

Métricas de desempenho

Acurácia

59%

Precisão

53%

Recall

80%

PREVISTO


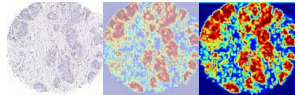
		cachorro	não é cachorro	
REAL	cachorro	40 VP	10 FN	acertos erros VP - Verdadeiros Positivos
	não é cachorro	35 FP	25 VN	VN - Verdadeiros Negativos FP - Falsos Positivos FN - Falsos Negativos

Quando realmente é um cachorro, o quão frequente o modelo classifica como cachorro?

$$\frac{40 \text{ (VP)}}{40 \text{ (VP)} + 10 \text{ (FN)}} = 0.80$$

Recall de 80%

Métricas de desempenho: Resumo

desempenho geral do modelo ▶	Acurácia	$\frac{VP + VN}{VP + VN + FP + FN}$	59%	
Quantos dos que o modelo classificou como cachorros são de fato cachorros? ▶	Precisão	$\frac{VP}{VP + FP}$	53%	
do total de verdadeiros, quantos conseguimos capturar? ▶	<i>Recall</i>	$\frac{VP}{VP + FN}$	80%	
média harmônica entre Precisão e Recall ▶	Medida F	$\frac{2 * Precisão * Recall}{Precisão + Recall}$	63%	

Valores categóricos x Valores contínuos

Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

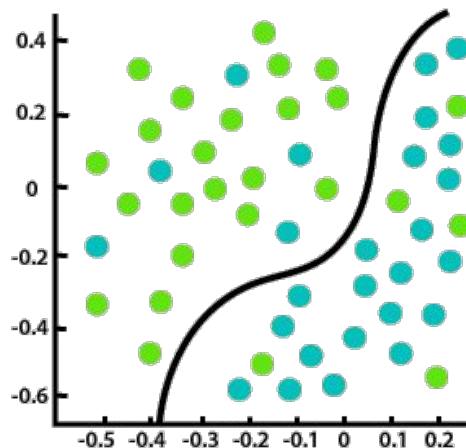
Motor	Emissão de CO2
2	196
2,4	221
1,5	136
3,5	255
3,5	244
3,5	230
3,7	232
3,7	255
3,7	267
2,4	212
2,4	225
3,5	239
5,9	359
5,9	359
4,7	338

[features] label/classe

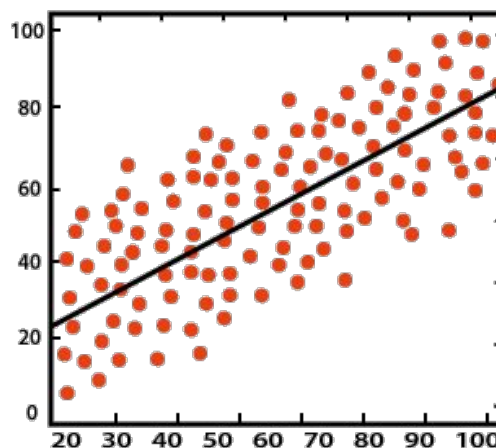
[independente] dependente

Classificação x Regressão

classificação de imagem
detecção de fraude
anti-spam
liberação de crédito
diagnóstico de doenças



Classification



Regression

previsão do tempo
valores de imóveis
preços de ações na bolsa

Técnicas de Classificação

CLASSIFICAÇÃO

- Métodos baseados em árvores de decisão
- Métodos baseados em regras
- Métodos baseados em memória
- Redes neurais
- Naive Bayes e Redes Bayesianas
- Máquinas de Vetores de Suporte

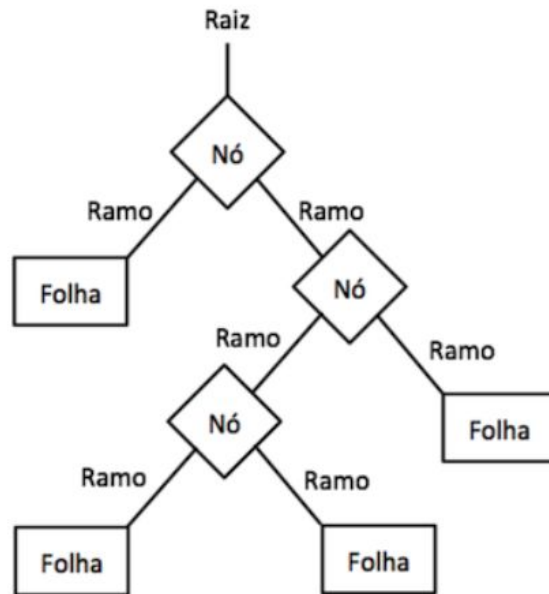
Técnicas de Classificação

CLASSIFICAÇÃO

- **Métodos baseados em árvores de decisão**
- Métodos baseados em regras
- Métodos baseados em memória
- Redes neurais
- Naive Bayes e Redes Bayesianas
- Máquinas de Vetores de Suporte

Árvore de decisão - Decision Tree

- A árvore é construída para tentar diminuir a incerteza
 - Quanto maior a incerteza, menor o ganho de informação
- Cada nó especifica o teste de algum atributo da instância
- Calcula-se o quanto cada feature organiza melhor os dados
 - Escolhe-se a melhor feature (com menor incerteza)



Entropia

- Medida da **aleatoriedade** ou **incerteza**
- Quanto menor a entropia, melhor a decisão (menos incertezas)
- **Ganho de informação**: Quanto menor a Entropia, maior o ganho de informação

$$H(S) = - \sum_{i=1}^n p_i \log_2 p_i$$

Árvore de decisão - Decision Tree

Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

[56 M ALTA BAIXA] ?

A, A, B, B, B, A, B, A, B, A

raiz

$$H(S) = - \sum_{i=1}^n p_i \log_2 p_i$$

$$p(A) = 5/10 = 0.5 \quad i = A \text{ e } B$$

$$p(B) = 5/10 = 0.5$$

$$\text{Entropia} = - (p(A) * \log_2 p(A)) - (p(B) * \log_2 p(B))$$

$$\text{Entropia} = - (0.5 * \log_2 0.5) - (0.5 * \log_2 0.5)$$

$$\text{Entropia} = - (0.5 * (-1)) - (0.5 * (-1))$$

$$\text{Entropia} = 0.5 + 0.5 = 1$$

Árvore de decisão - Decision Tree

Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

[56 M ALTA BAIXA] ?



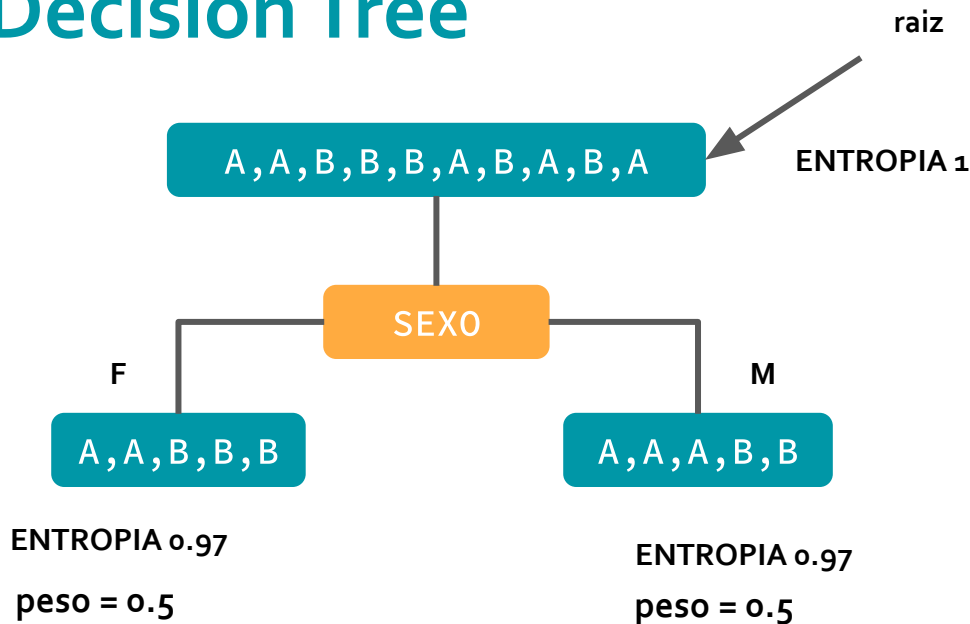
- Próximo nó da árvore? (feature):

A com maior ganho de informação!

Árvore de decisão - Decision Tree

Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

[56 M ALTA BAIXA] ?



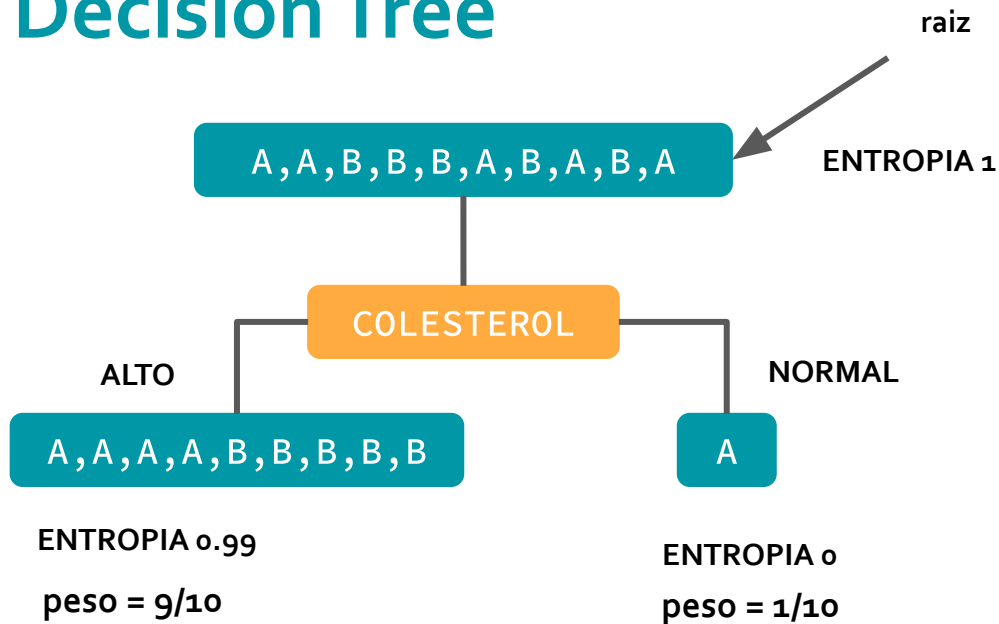
$$\text{ganho} = 1 - ((0.5 * 0.97) + (0.5 * 0.97)) = 0.03$$

$$\text{ganho} = Entropia(\text{pai}) - \sum \text{peso}(\text{filhos}) * Entropia(\text{filhos})$$

Árvore de decisão - Decision Tree

Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

[56 M ALTA BAIXA] ?



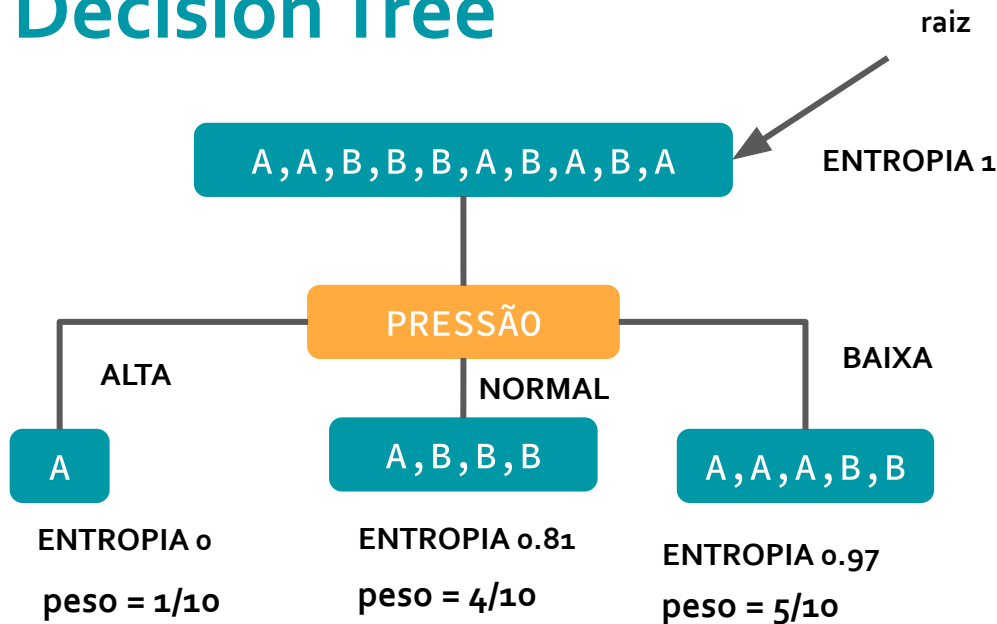
$$\text{ganho} = 1 - ((0.99 * 0.9) + (0 * 0.1)) = 0.11$$

$$\text{ganho} = Entropia(\text{pai}) - \sum \text{peso}(\text{filhos}) * Entropia(\text{filhos})$$

Árvore de decisão - Decision Tree

	0.03	0.19	0.11	
Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

[56 M ALTA BAIXA] ?



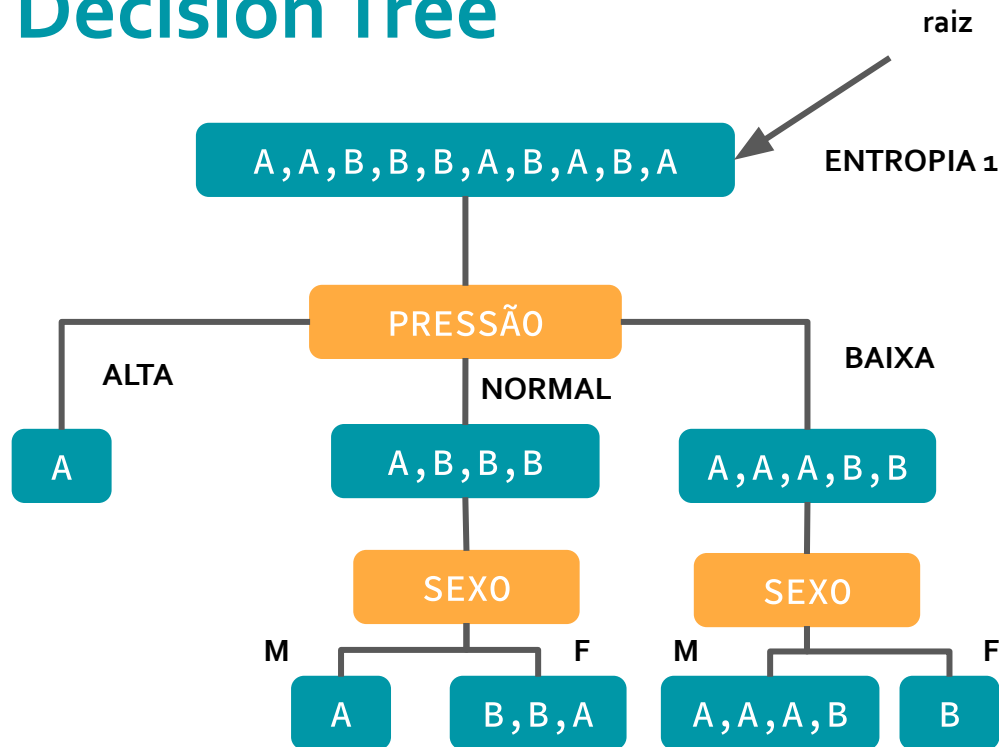
$$\text{ganho} = 1 - ((0 * 0.1) + (0.81 * 0.4) + (0.97 * 0.5)) = 0.19$$

$$\text{ganho} = Entropia(\text{pai}) - \sum \text{peso}(\text{filhos}) * Entropia(\text{filhos})$$

Árvore de decisão - Decision Tree

Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

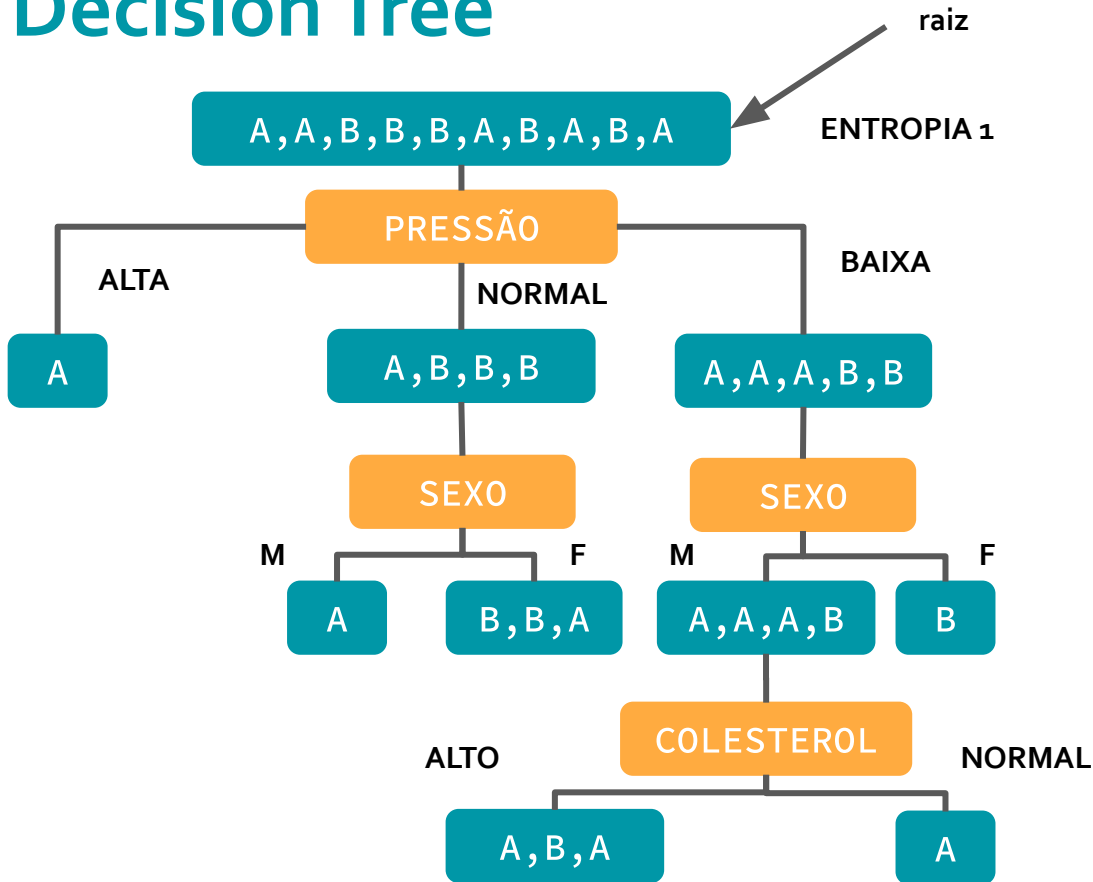
[56 M ALTA BAIXA] ?



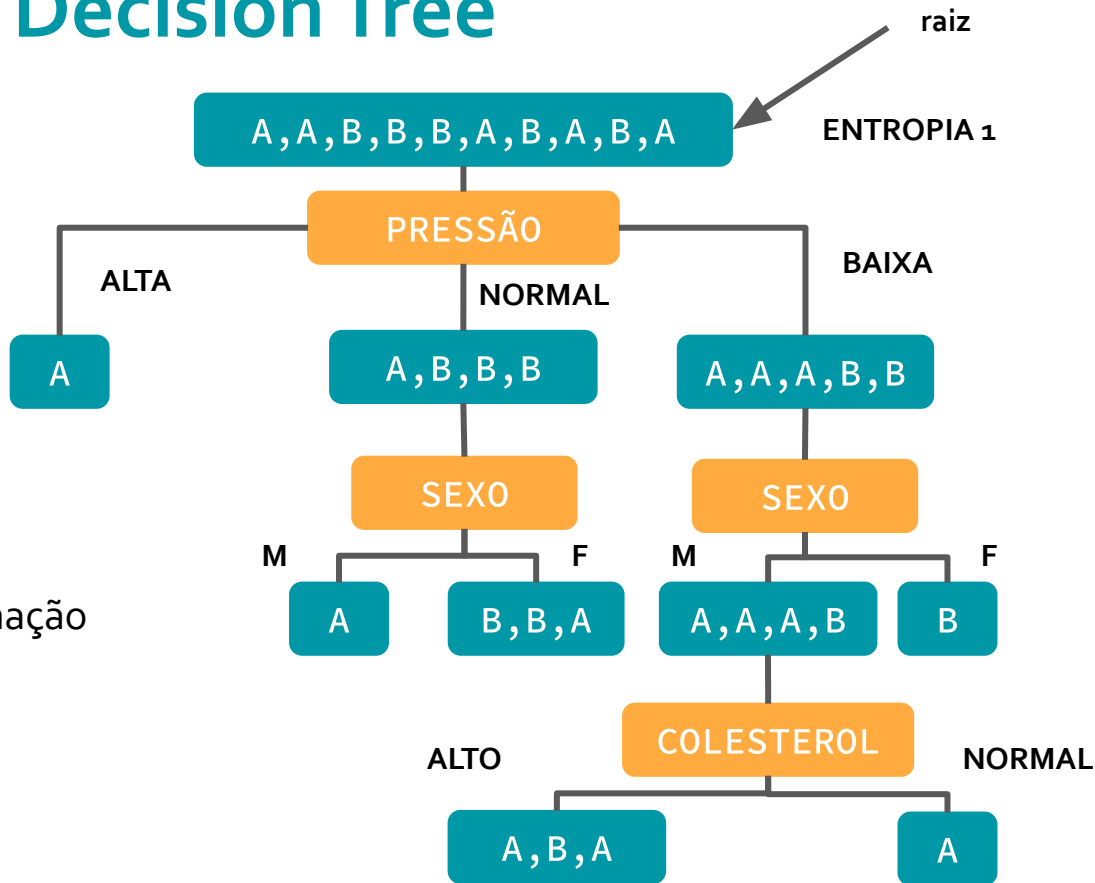
Árvore de decisão - Decision Tree

Idade	Sexo	Pressão Arterial	Colesterol	Droga
23	F	ALTA	ALTA	A
47	M	BAIXA	ALTA	A
47	M	BAIXA	ALTA	B
28	F	NORMAL	ALTA	B
61	F	BAIXA	ALTA	B
22	F	NORMAL	ALTA	A
49	F	NORMAL	ALTA	B
41	M	BAIXA	ALTA	A
60	M	NORMAL	ALTA	B
43	M	BAIXA	NORMAL	A

[56 M ALTA BAIXA] ?



Árvore de decisão - Decision Tree



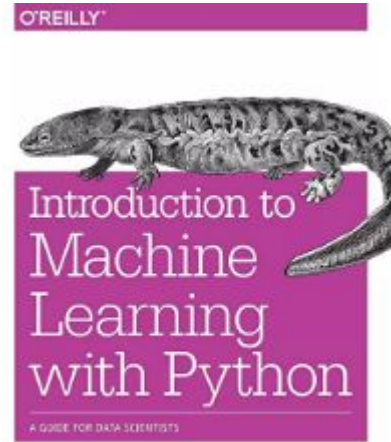
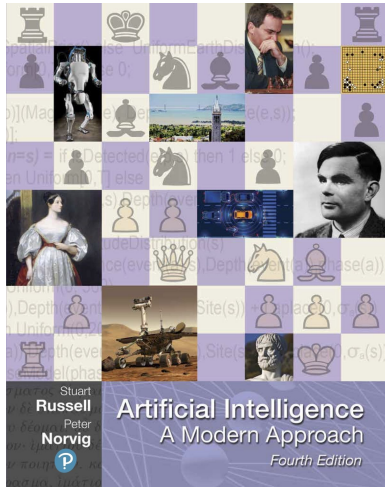
■ Quando parar?

1. Até não ter mais ganho de informação
2. Especificar a altura da árvore

Parte prática 2



Livros e links!



Andreas C. Müller & Sarah Guido

- [Curso IA para todos do prof. Diogo Cortiz da PUC-SP](#)
- [Livros gratuitos de Ciência de Dados](#)
- [Materiais do prof. Regis da UFC](#)

