ELDORADO

Trilha ciência de dados com Python

Aula 14

TIC em trilhas

Chamada

TIC em trilhas

Faísca

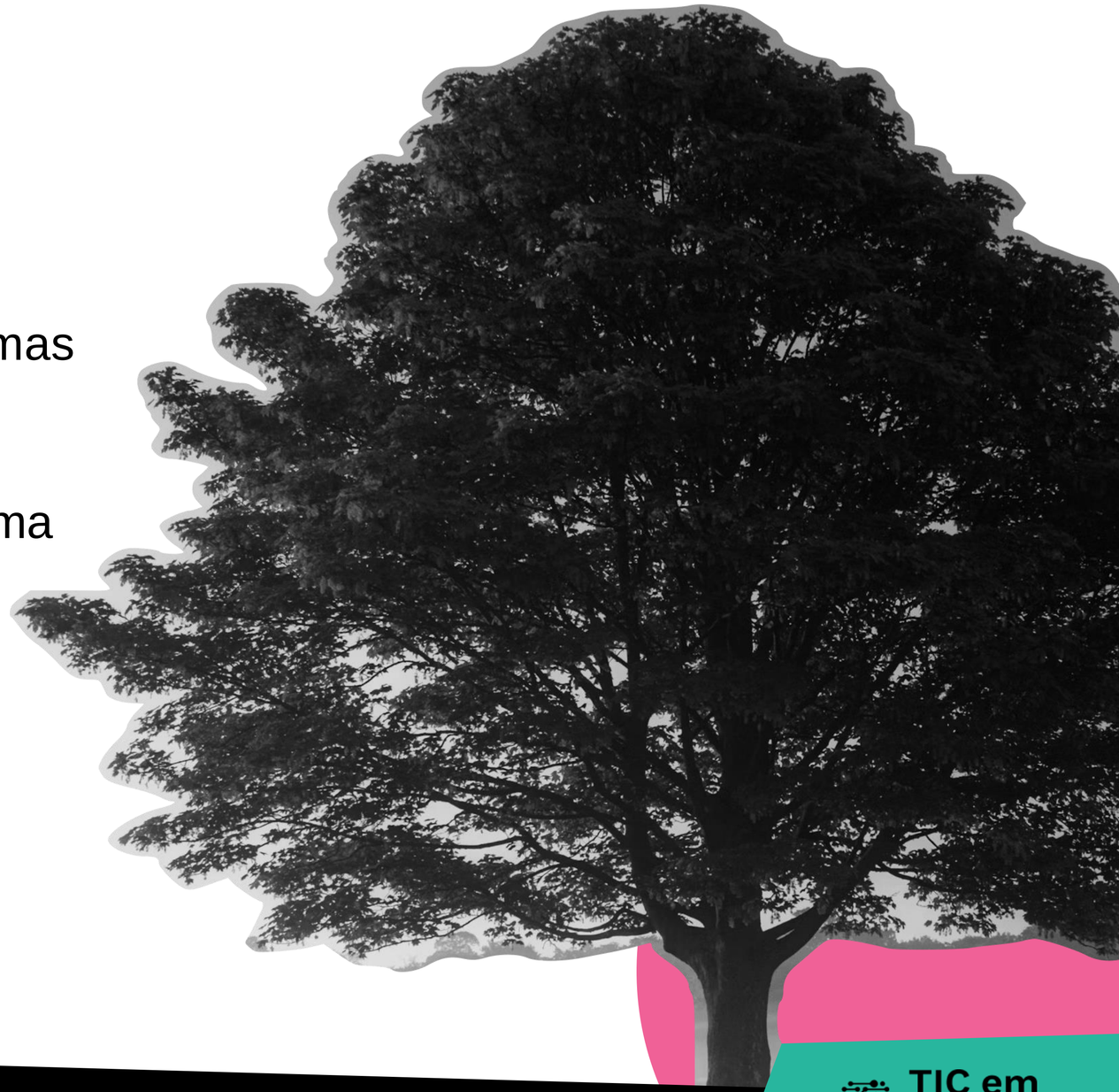
TIC em trilhas

Roteiro de hoje!

- Definição
- Exemplo de árvore de decisão
- Geração da árvore de decisão
- Métricas utilizadas para selecionar a melhor divisão
- Avaliação do desempenho em Árvore de Decisão
- Atividade

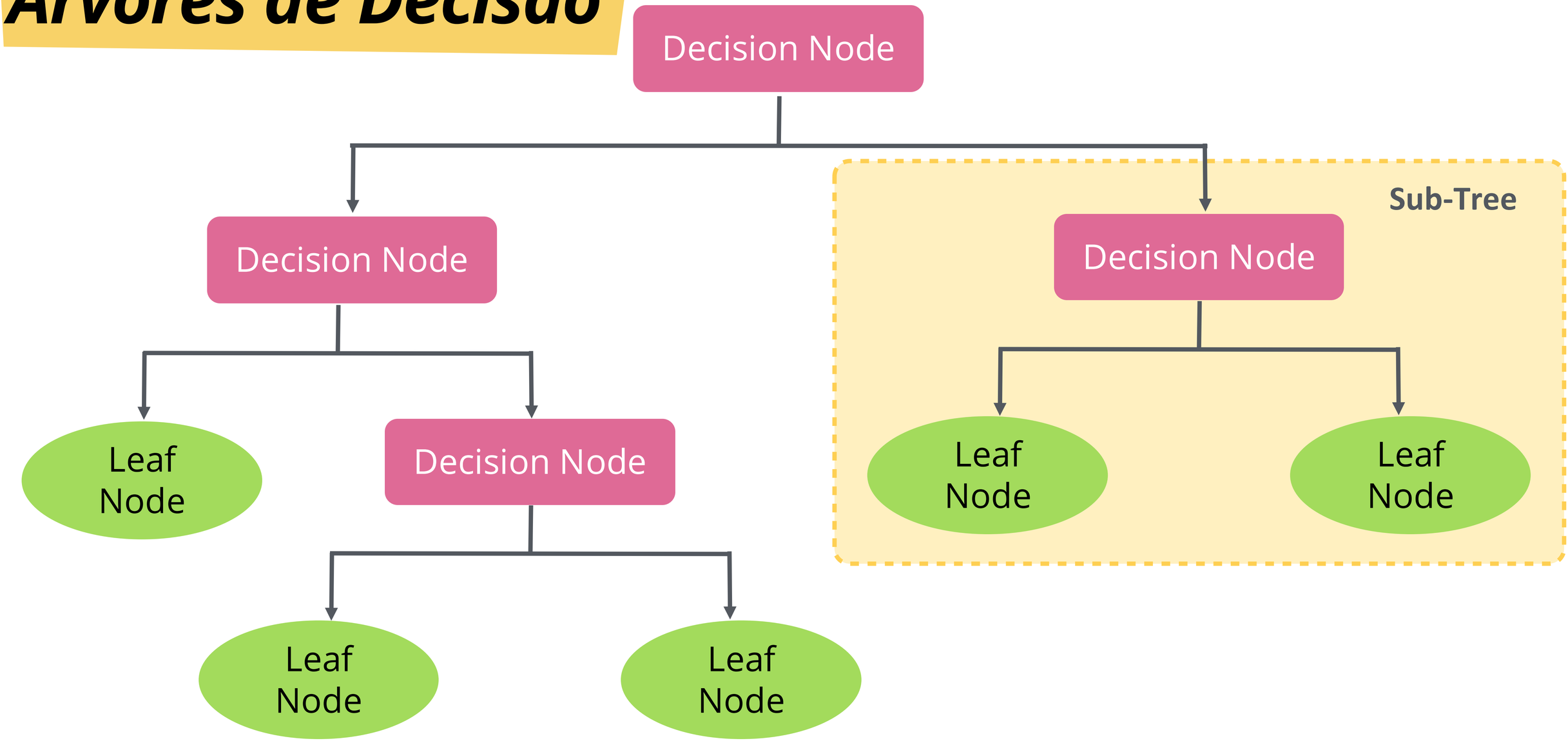
Árvores de Decisão

- **Definição**
  - Utiliza a estratégia dividir para conquistar.
  - Um problema complexo é dividido em problemas mais simples (subproblemas).
  - Para cada subproblema é aplicada uma mesma estratégia recursivamente
- **Algortimos**
  - ID3 (Quilan, 1979).
  - **CART (Breiman et al., 1984).**
  - C4.5 (J48 no Weka) (Quilan, 1993) .



TIC em trilhas

Árvores de Decisão



Akinator

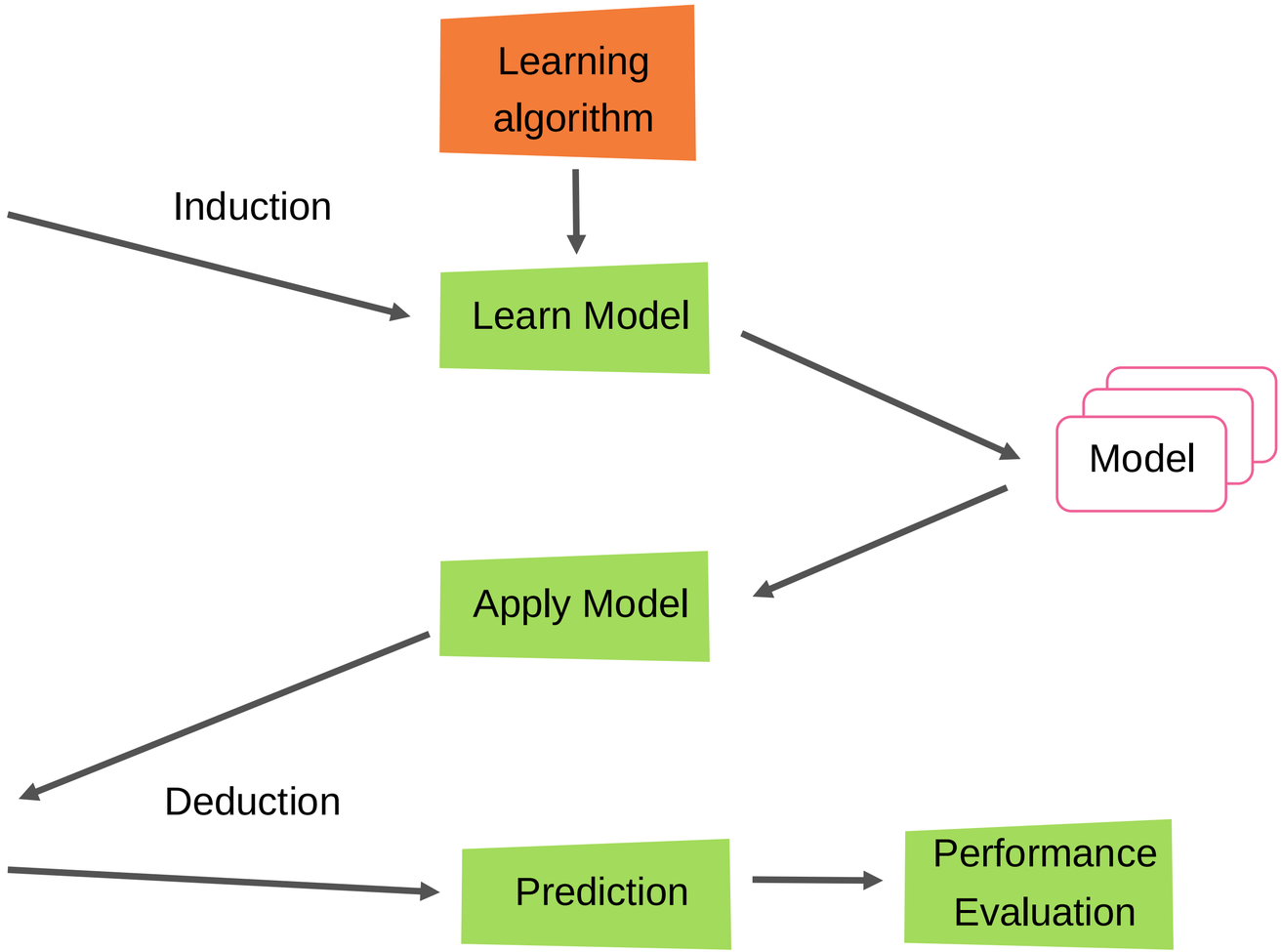
Esquema da Tarefa de Classificação

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Set

TID	Attrib 1	Attrib2	Attrib3	Class
11	No	Small	55k	?
12	Yes	Medium	80k	?
13	Yes	Large	95k	?
14	No	Small	95k	?
15	No	Large	67k	?

Test Set



Exemplos de Árvore de Decisão

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

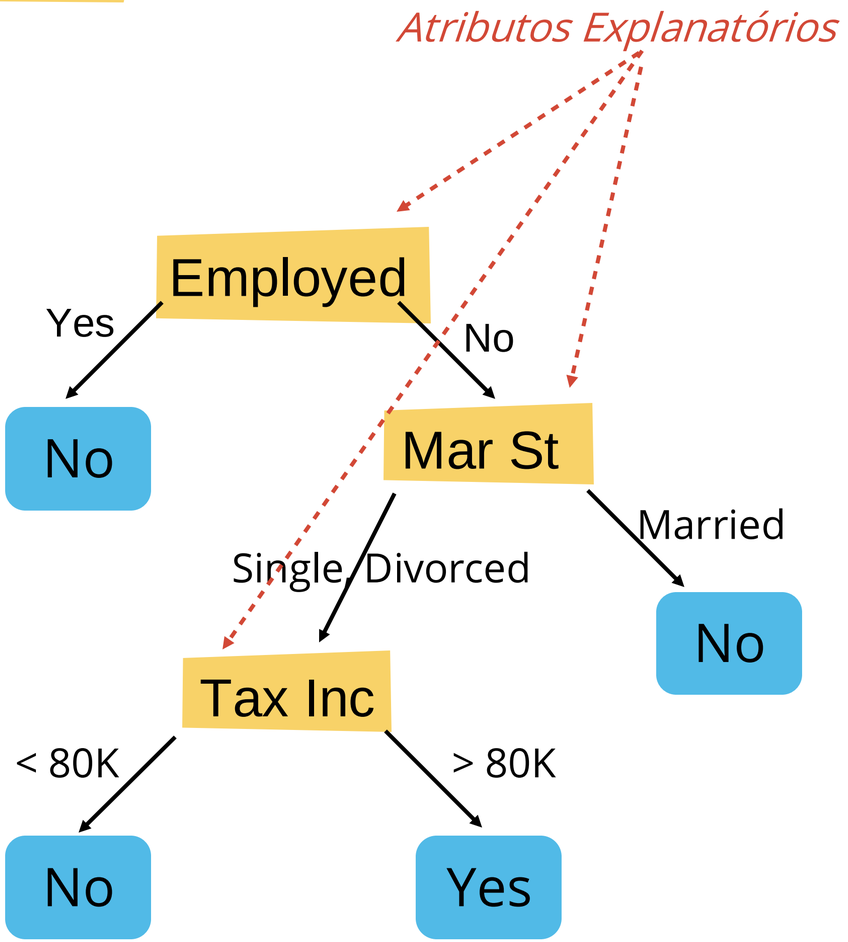
categórico

categórico

contínuo

classe

Training Set



Modelo: Árvore de Decisão

Exemplos de Árvore de Decisão

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

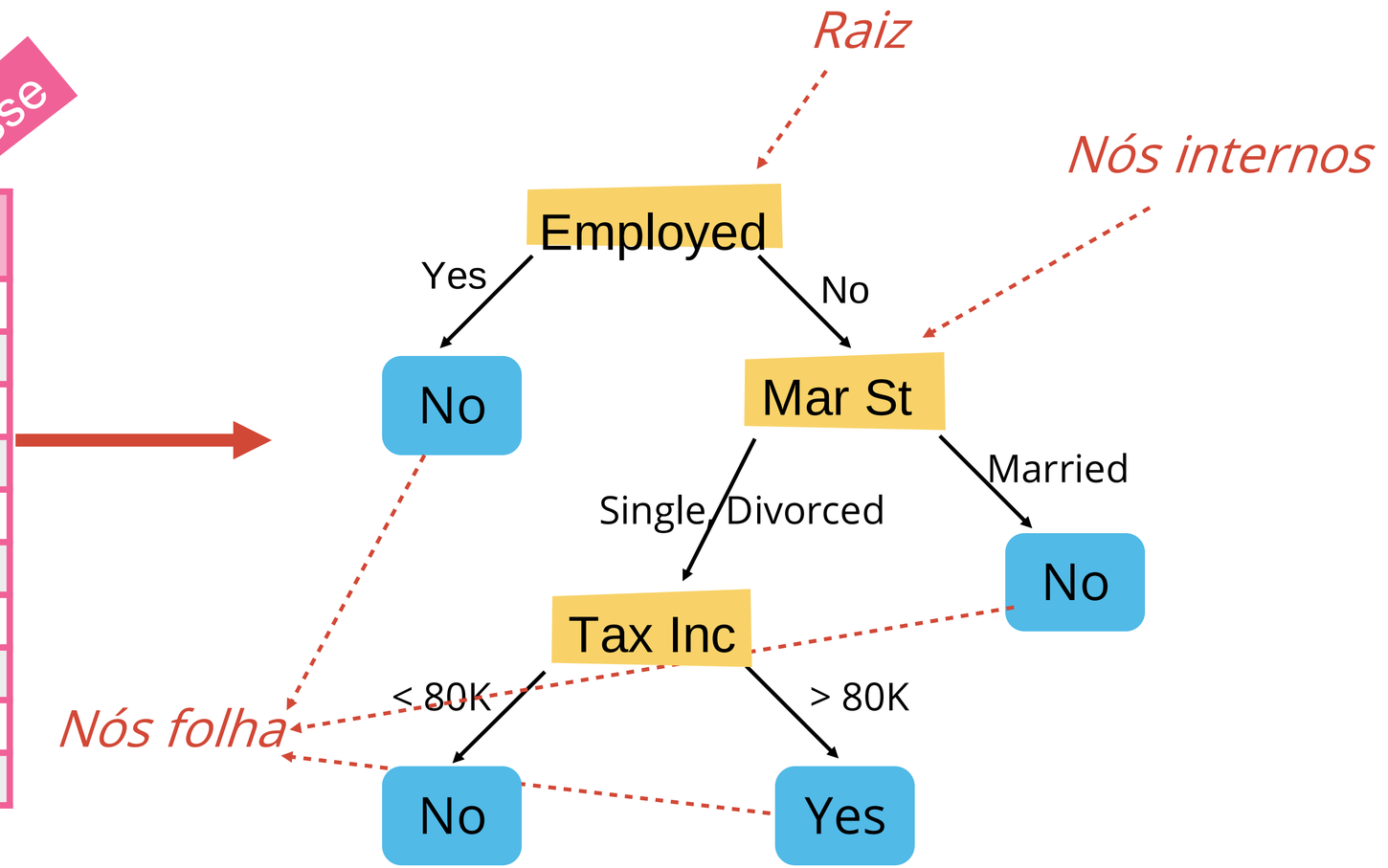
categórico

categórico

contínuo

classe

Training Set



Modelo: Árvore de Decisão

Exemplos de Árvore de Decisão

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

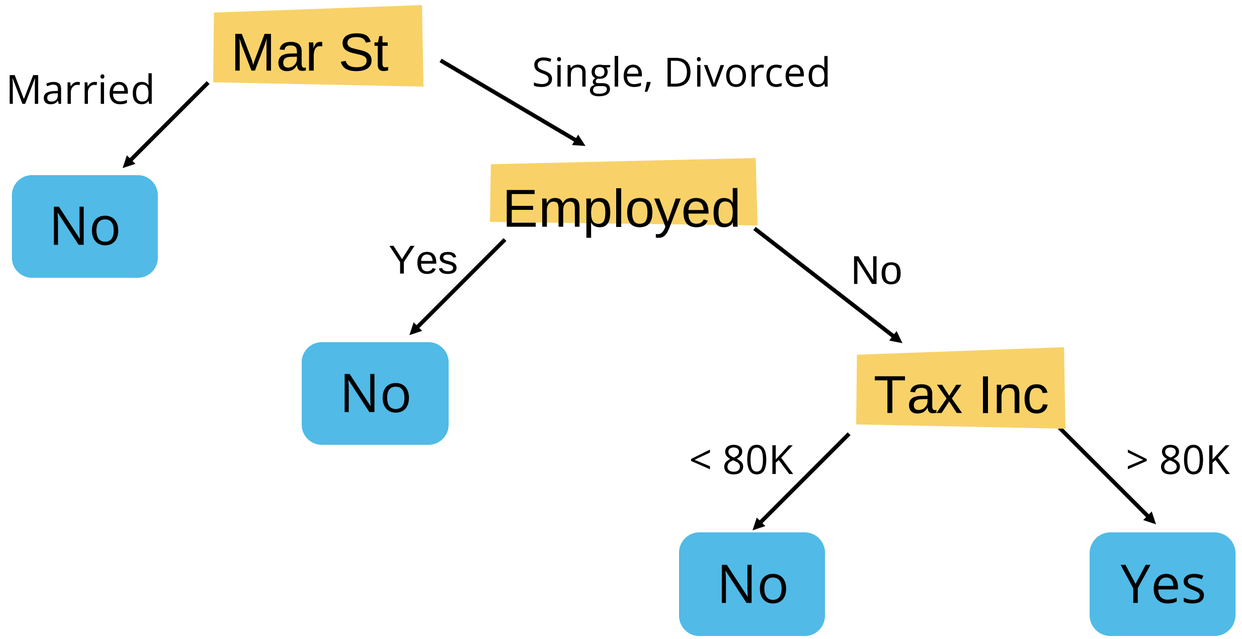
categórico

categórico

contínuo

classe

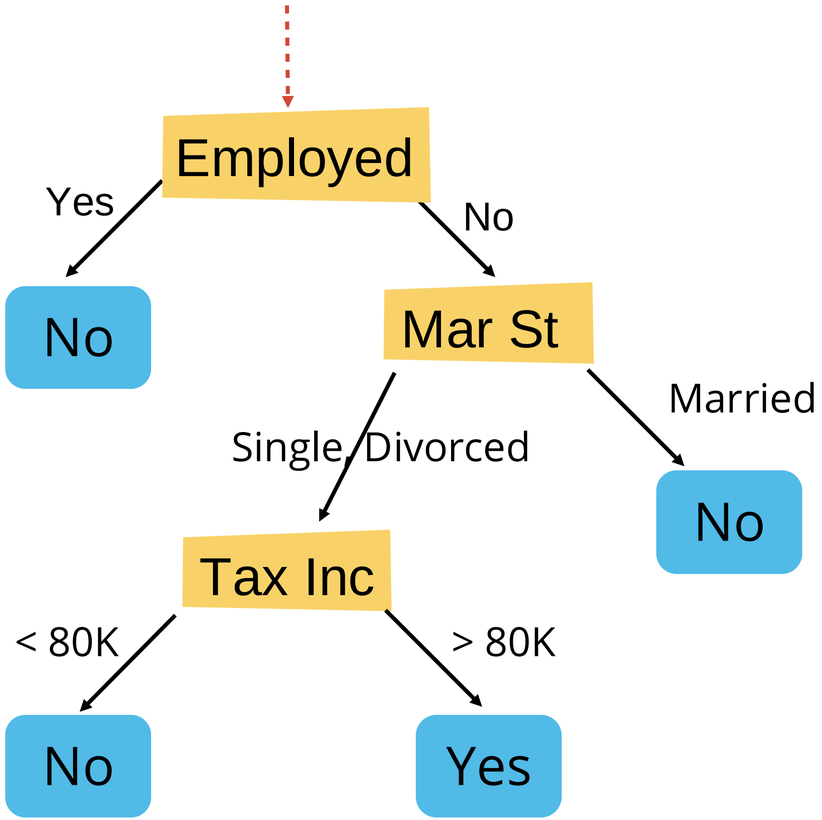
Training Set



Pode existir mais de uma árvore de decisão adequada para os mesmos dados!

Aplicando o Modelo aos Dados de Teste

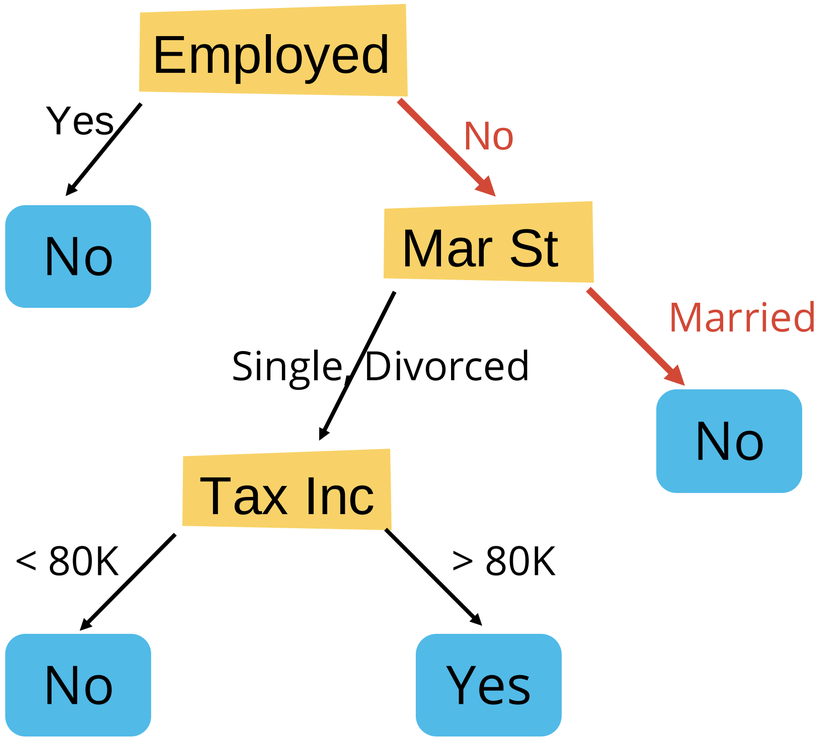
Início na raiz da árvore.



Training Set

Employed	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Aplicando o Modelo aos Dados de Teste



Training Set

Employed	Marital Status	Taxable Income	Cheat
No	Married	80K	?

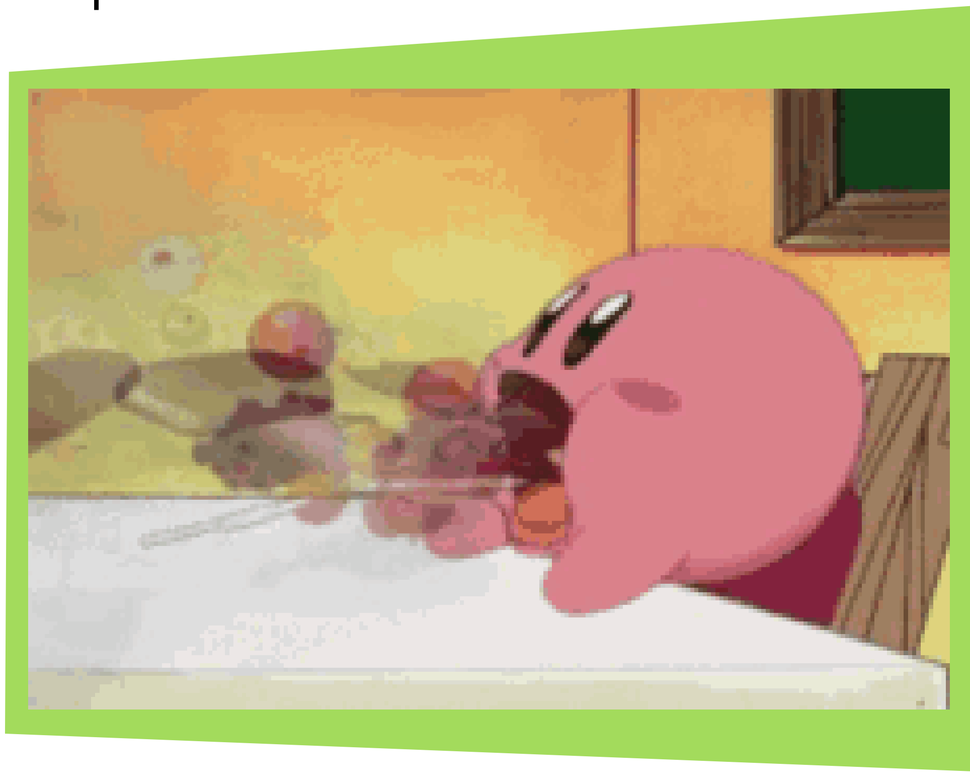
Atribuir “NO” para Calote.

Indução de Árvores de Decisão

- **Características**
  - Pequenas árvores de decisão são muito fáceis de **interpretar**.
  - A **construção** de árvores é computacionalmente **barata** mesmo para uma grande quantidade de dados.
  - A **classificação** dos dados de testes em uma árvore de decisão é extremamente **rápida**.
  - Os algoritmos de árvores de decisão são bastante **robustos** para a presença de ruídos, especialmente quando possuem métodos para evitar o *overfitting*.
  - A presença de atributos **redundantes** afeta negativamente a acurácia de árvores de decisão.

Indução de Árvores de Decisão

- **Estratégia Gulosa (Greedy).**
  - Particionar os registros baseado no teste de um atributo que otimiza um certo critério.
- Problemas:
  - Determinar como particionar os registros.
  - Como especificar a condição de teste para o atributo?
  - Como determinar qual é o melhor particionamento?
- Determinar quando parar de particionar.



Indução de Árvores de Decisão

- Depende do tipo do atributo
  - Nominal
    - cor, identificação, profissão, ...
  - Ordinal
    - gosto (ruim, médio, bom), dias da semana, ...
  - Contínuo (numérico)
    - peso, tamanho, idade, temperatura, ...
- Depende do número de ramos para particionar
  - Particionamento em 2 ramos.
  - Particionamento em n ramos.

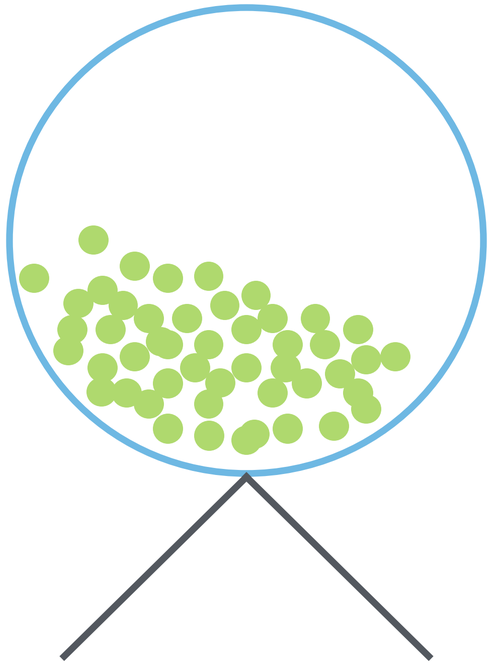
Particionamento em Atributos Contínuos

- Diferentes maneiras de tratar:
  - Discretização para transformar em um atributo categórico ordinal.
    - Estático – discretizado uma vez no início.
    - Dinâmico – intervalos podem ser achados por particionamento em intervalos iguais, em frequências iguais, ou agrupamento.
  - Teste Binário: $(A < v)$ ou $(A \geq v)$
    - Considera todos os possíveis pontos de corte e procura o melhor.

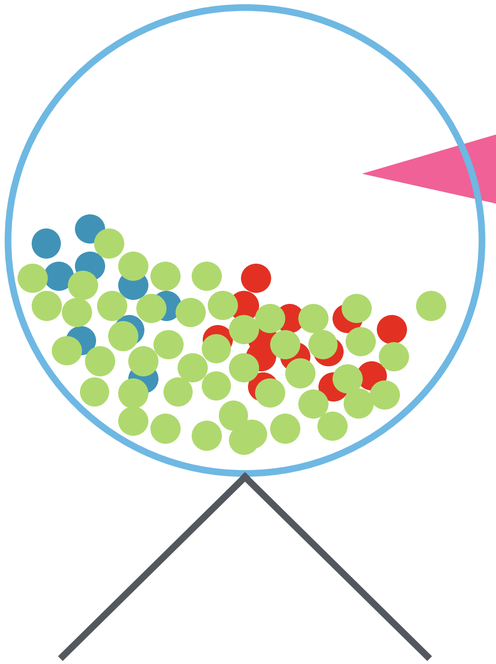
Para Determinar o Melhor Ponto de Particionamento

- **Grau de impureza**

Totalmente puro



Maior grau de impureza



TIC em trilhas

Para Determinar o Melhor Ponto de Particionamento

- **Abordagem Gulosa**:
  - Nodos com distribuição homogênea de classes são preferidos.
- Necessita de uma métrica para medir a impureza do nodo:

C0: 5 C1: 5	C0: 9 C1: 1
Não-homogêneo Alto grau de impureza	Homogêneo Baixo grau de impureza

Classificação

Métricas Utilizadas para Selecionar a Melhor Divisão

TIC em trilhas

Métricas para Avaliar a Impureza de Nodos

- **Índice Gini**

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

- Entropia

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

- Erros de classificação

$$Error(t) = 1 - \max_i P(i \mid t)$$

TIC em trilhas

Métricas de Impureza: GINI

- Índice Gini para um dado nodo t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTA: *p(j / t)* é a frequência relativa da classe j no nodo t).

Máximo (1 - 1/n_c) quando registros são igualmente distribuidos entre todas as classes, implicando na informação menos interessante.

Mínimo (0.0) quando todos os registros pertencem a uma única classe, implicando na informação mais interessante.

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

Particionamento baseado no Índice GINI

- Usado pelos algoritmos CART, SLIQ, SPRINT.
- Quando um nodo p é particionado em k partições (filhos), a qualidade do particionamento é calculado por,

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

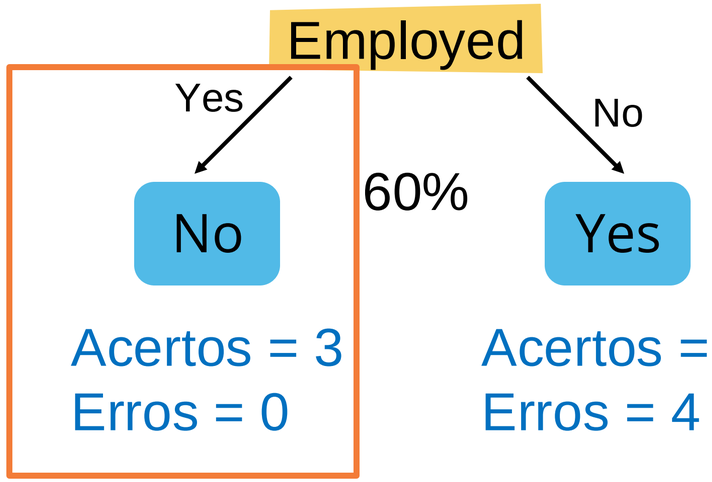
onde, n_i = número de registros no filho i,
n = número de registros no nodo p.

Atributos Categóricos: Calculando o GINI

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

categórico categórico contínuo classe

Training Set



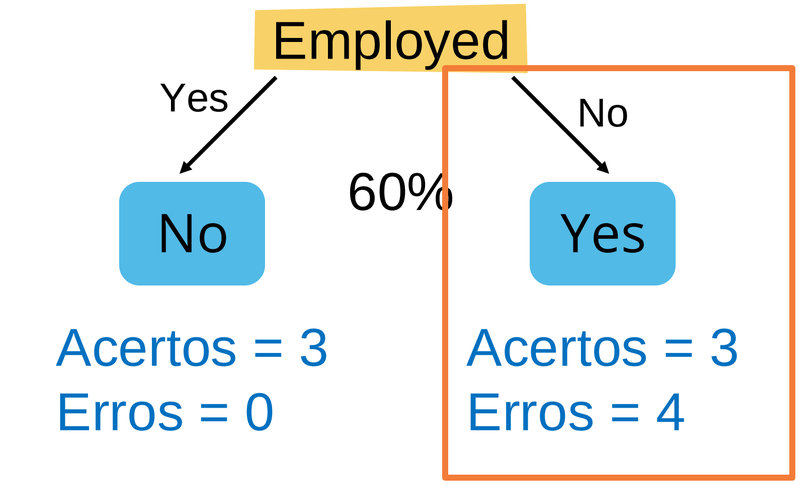
$Gini = 1 – (3/3)^2 – (0/3)^2$
$Gini = 1 – 1 – 0$
$Gini = 0,0$

$Gini = 1 – (3/7)^2 – (4/7)^2$
$Gini = 1 – 9/49 – 16/49$
$Gini = (49 – 9 – 16)/49$
$Gini = 0,49$

$Gini_{split} = (3/10)*0,0 + (7/10)*0,49$
$Gini_{split} = 0 + 0,34$
$Gini_{split} = 0,34$

Atributos Categóricos: Calculando o GINI

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



$$Gini = 1 - (3/3)^2 - (0/3)^2$$
$$Gini = 1 - 1 - 0$$
$$Gini = 0,0$$

$$Gini = 1 - (3/7)^2 - (4/7)^2$$
$$Gini = 1 - 9/49 - 16/49$$
$$Gini = (49 - 9 - 16)/49$$
$$Gini = 0,49$$

$$Gini_{split} = (3/10)*0,0 + (7/10)*0,49$$
$$Gini_{split} = 0 + 0,34$$
$$Gini_{split} = 0,34$$

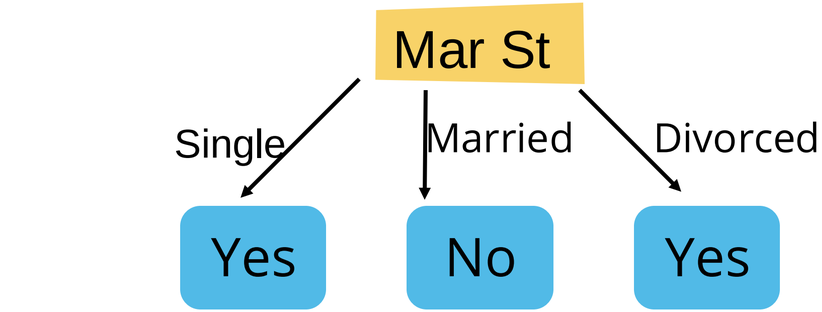
TIC em trilhas

Atributos Categóricos: Calculando o GINI

categórico | categórico | contínuo | classe

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Set



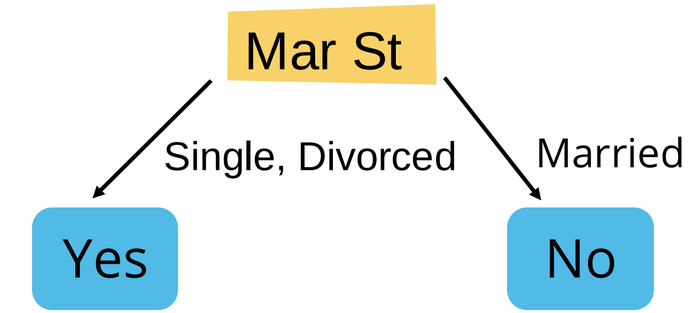
Acertos = 2
Erros = 2

Acertos = 4
Erros = 0

Acertos = 1
Erros = 1

70%

$Gini_{split}$ = 0,3



Acertos = 3
Erros = 3

Acertos = 4
Erros = 0

70%

$Gini_{split}$ = 0,3

TIC em trilhas

Atributos Contínuos: Calculando Índice GINI

- Classificar valores existentes.
- Pesquisar linearmente estes valores, apurando a população envolvida, e calculando o índice GINI.
- Escolher a posição de particionamento que apresenta o menor índice GINI.

Calote	N		N		N		S		S		S		N		N		N		N	
	Rendim. Tributáveis																			
Valores Ordenados →	60		70		75		85		90		95		100		120		125		220	
Posições de particionamento →	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
S	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
N	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		***0.300***		0.343		0.375		0.400		0.420	

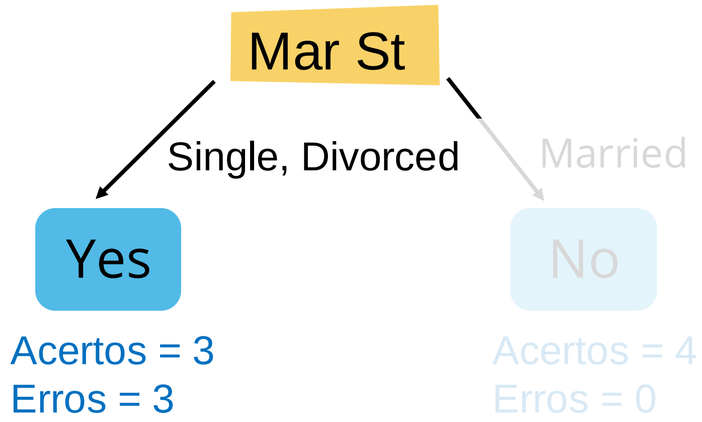
Induzindo o 2º Nível da Árvore de Decisão

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

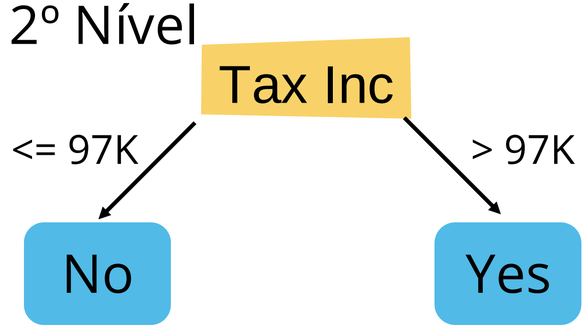
categórico | categórico | contínuo | classe

Training Set

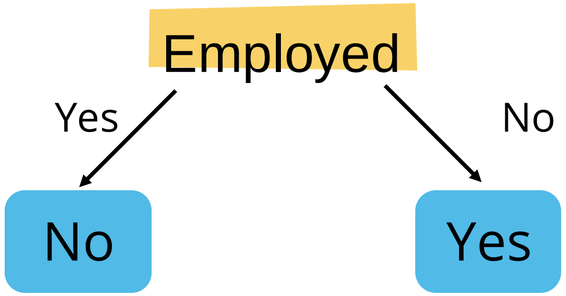
1º Nível



2º Nível



$Gini_{split} = 0,25$



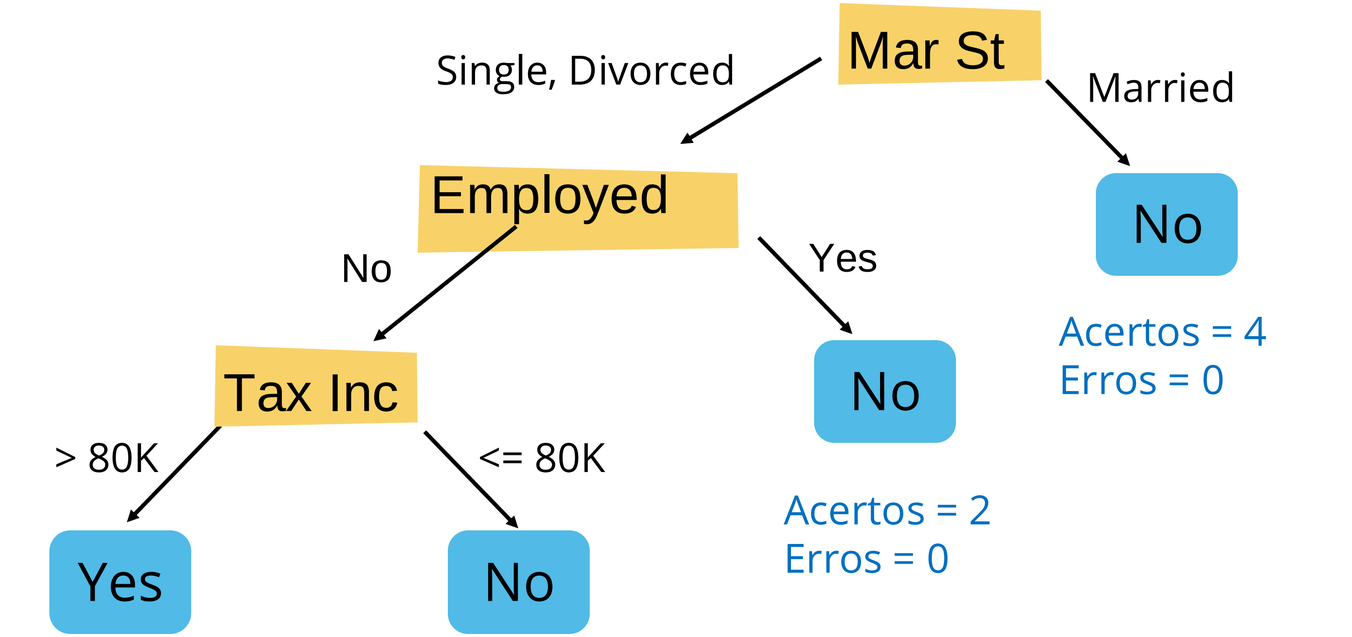
$Gini_{split} = 0,25$

Induzindo o 2º Nível da Árvore de Decisão

TID	Employed	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

categórico | categórico | contínuo | classe

Training Set

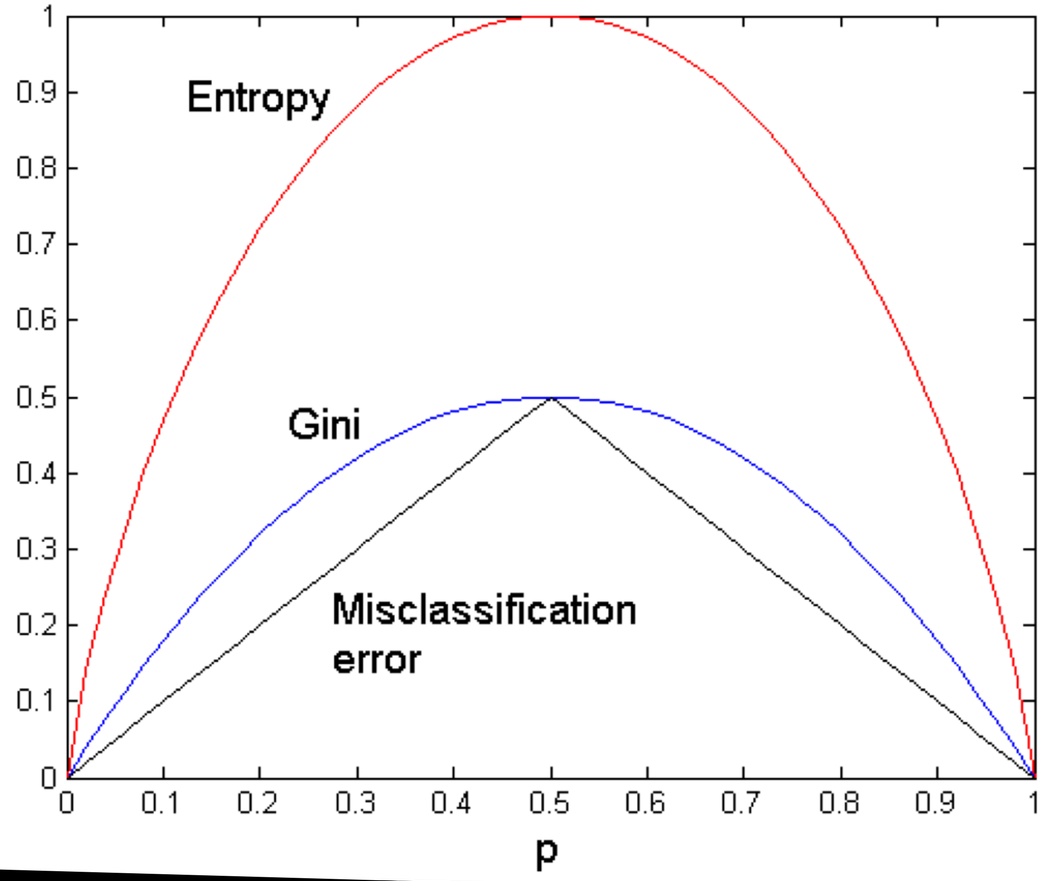


$Gini_{split} = 0,0$

Medidas para Selecionar a Melhor Divisão

- As medidas são baseadas no grau de impureza dos nodos filhos.
- Quanto menor o grau de impureza mais distorcida será a distribuição da classe.
- Por exemplo:
  - Um nodo com classe de distribuição uniforme (0,1) tem impureza zero.
  - Um nodo com distribuição de classe uniforme (0.5,0.5) possui uma impureza mais alta.

Comparação entre as medidas de impurezas para problemas de classificação binária



TIC em trilhas

Medidas para Selecionar a Melhor Divisão

- Determine o Gini, a Entropia e o Erro dos nodos abaixo.

Nodo N_1	Quant
Classe=0	0
Classe=1	6

Gini =

Nodo N_1	Quant
Classe=0	1
Classe=1	5

Gini =

Nodo N_1	Quant
Classe=0	3
Classe=1	3

Gini =

Árvores de Decisão

- **Vantagens**
  - Simples de visualizar e entender
  - Não necessita muita preparação para os dados (pre-processamento), tais como normalização
    - Apenas não aceita valores faltantes
  - O custo é logaritmo a quantidade de dados usados para treinar a árvore
  - Suporta dados numéricos e categóricos.
  - Modelo caixa branca: fácil interpretação
  - Possível reproduzir o modelo utilizando testes estatísticos

Árvores de Decisão

- **Desvantagens**
  - Indutores de árvores de decisão podem criar modelos muito complexos que não generalizam bem todos os dados (overfitting)
    - Para evitar esse problema deve ser definido um número mínimo de objetos nos nodos folhas ou um número máximo de profundidade da árvore
  - Pequenas variações no dataset podem gerar modelos instáveis
    - Esse problema pode ser atenuado usando árvore de decisão em conjuntos menores.
  - Podem criar modelos tendenciosos a classes dominantes (**bias**)
    - Recomenda-se equilibrar o conjunto de dados ante de ajustar a árvore de decisão.

Classificação

Avaliando o Desempenho de um Classificador

TIC em trilhas

Avaliação de Desempenho

- Matriz de Confusão:

	CLASSE PREVISTA		
CLASSE REAL		Classe=SIM	Classe=NAO
	Classe=SIM	a (TP)	b (FN)
	Classe=NAO	c (FP)	d (TN)

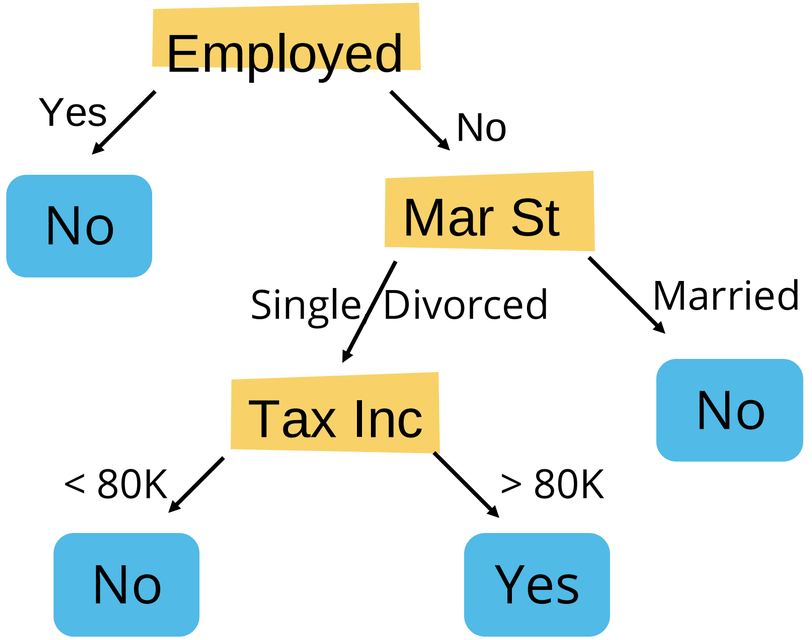
a: **TP** (true positive)
verdadeiro positivo
b: **FN** (false negative)
falso negativo

c: **FP** (false positive)
falso positivo
d: **TN** (true negative)
verdadeiro negativo

Métricas para Avaliação de Desempenho

TID	Employed	Marital Status	Taxable Income	Cheat
	categórico	categórico	contínuo	classe
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Set



		CLASSE PREVISTA	
		Classe=SIM	Classe=NAO
CLASSE REAL	Classe=SIM	3	0
	Classe=NAO	0	7

Acurácia = 100%

TIC em trilhas

Métricas para Avaliação de Desempenho

	CLASSE PREVISTA		
CLASSE REAL		Classe=SIM	Classe=NAO
	Classe=SIM	**3 (TP)**	**0 (FN)**
	Classe=NAO	**4 (FP)**	**3 (TN)**

- PAcuracy: (TP+TN)/(TP+FN+FP+TN) = 60%
  - Percentual de acertos.
- Recall (sensibilidade): TP/(TP+FN) = 100%
  - Representa as instâncias que deveriam ser da classe **S** mas foram classificadas na classe **N.** Mais direcionado para a classe real.

Métricas para Avaliação de Desempenho

	CLASSE PREVISTA		
CLASSE REAL		Classe=SIM	Classe=NAO
	Classe=SIM	**3 (TP)**	**0 (FN)**
	Classe=NAO	**4 (FP)**	**3 (TN)**

- Precision (especificidade): TP/(TP+FP) = 43%
  - Representa as instâncias que deveriam ser da classe **N** mas foram classificadas na classe **S.**

    Direcionado para a classe prevista
- F1-Score: (2×(Recall×Precision))/(Recall+Precision) = 60,14%
  - Equilíbrio entre Precision e Recall
  - Representa a distribuição de classe desigual

Métricas para Avaliação de Desempenho

- Acuracy:
  - Percentual de acertos.
  - (TP+TN)/(TP+FN+FP+TN) = 60%

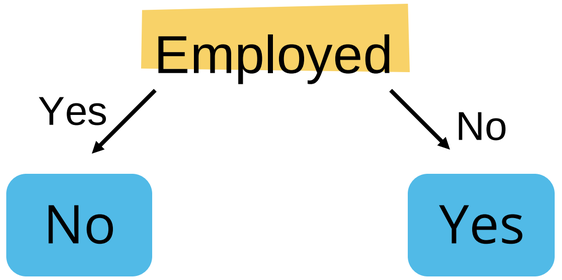
	CLASSE PREVISTA		
CLASSE REAL		Classe=SIM	Classe=NAO
	Classe=SIM	**3 (TP)**	**0 (FN)**
	Classe=NAO	**4 (FP)**	**3 (TN)**

- Recall (sensibilidade):
  - TP/(TP+FN) = 100%
- Precision (especificidade):
  - TP/(TP+FP) = 43%
- F1-Score: (2×(Recall×Precision))/(Recall+Precision) = 60,14%

Métricas para Avaliação de Desempenho

TID	Employed	Marital Status	Taxable Income	Cheat
	categórico	categórico	contínuo	classe
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Set



	CLASSE PREVISTA		
CLASSE REAL		Classe=SIM	Classe=NAO
	Classe=SIM	**3**	**0**
	Classe=NAO	**4**	**3**

Acurácia = 60%

Sensibilidade (*Recall*) = 100%
(percentual de positivos pegos)

Especificidade (*Precision*) = 43%
(percentual de negativos pegos)

Exercício

- Utilize o **dataset** zoo_2.csv e execute o algoritmo utilizando o método holdout, reservando apenas 20% dos dados para teste.
- Gere a matriz de confusão para os datasets de treino e teste.
- Compute as métricas (precision, recall e f1-score) para os conjutos de treino e teste.
- ***Avalie*** e ***compare*** os resultados obtidos pelo dataset de treino e teste e identifique se houve *overfitting*, *underfitting* ou se o modelo induzido gerado é adequado para utilizar em dados não vistos.
- Caso os resultados não estejam bons, altere os parâmetros do algoritmo para tentar melhorar o desempenho do modelo.

Créditos

- Adaptação dos slides de Pang-Ning Tan
  - Michigan State University
  - http://www.cse.msu.edu/~ptan/
  - ptan@cse.msu.edu
- Adaptação dos slides de Eamon Keogh
  - University of California at Riverside
  - http://www.cs.ucr.edu/~eamonn/
  - eamonn@cs.ucr.edu
- Adaptação dos slides de Ricardo Campello e Eduardo Hruschka
  - Universidade de São Paulo (ICMC)
- Adaptação dos slides de Rodrigo Barros
  - Pontifícia Universidade Católica do Rio Grande do Sul (PPGCC)

Referências

- Breiman, L., Freidman, J., Olshen, R. e Stone, C. (1984). *Classification and Regression Trees.* Wadsworth International Group., USA.
- Faceli, K.; Lorena, A.C.; Gama, J.; de Carvalho, A.C.P.L.F. Inteligência Artificial: Uma abordagem de aprendizado de máquina. LTC, Rio de Janeiro, 2011.
- Quilan, R. (1979). *Discovering rules by induction from large collections of examples.* In: Michie, D. (Ed.) Expert Systems in the Microelectronic Age, p. 168-201. Edinburgh University Press.
- Quilan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA.
- TAN, P-N; STEINBACH, M.; KUMAR,V. *Introduction to Data Mining.* Pearson, 2006.

Referências

- https://medium.com/data-hackers/%C3%A1rvore-de-decis%C3%A3o-88c7d0fd7a31
- https://scikit-learn.org/stable/modules/tree.html

ELDORADO

Trilha ciência de dados com Python

Aula 14

TIC em trilhas