ELDORADO

Trilha Ciência de dados com Python

Aula 15

TIC em trilhas

Chamada

TIC em trilhas

Faísca

TIC em trilhas

Roteiro de hoje!

- Definição
- Representação
- Funcionamento
- Atividade

Random Forests

- **Definição**
  - Combina simplicidade das árvores de decisão com a flexibilidade e aleatoriedade para melhorar a precisão.
  - Constrói conjuntos (ensembles) de árvores de decisão, melhorando o desempenho de generalização, manipulando instâncias e atributos de entrada.
  - Utiliza uma amostra de boostrap diferente de treinamento de dados para aprender a partir de árvores de decisões.
  - Para cada nodo, o melhor critério de divisão é escolhido entre um pequeno conjunto de atributos selecionados aleatoriamente. (BOOTSTRAP)

Random Forests - Representação

Training dataset

Bootstrap

Ensemble of trees

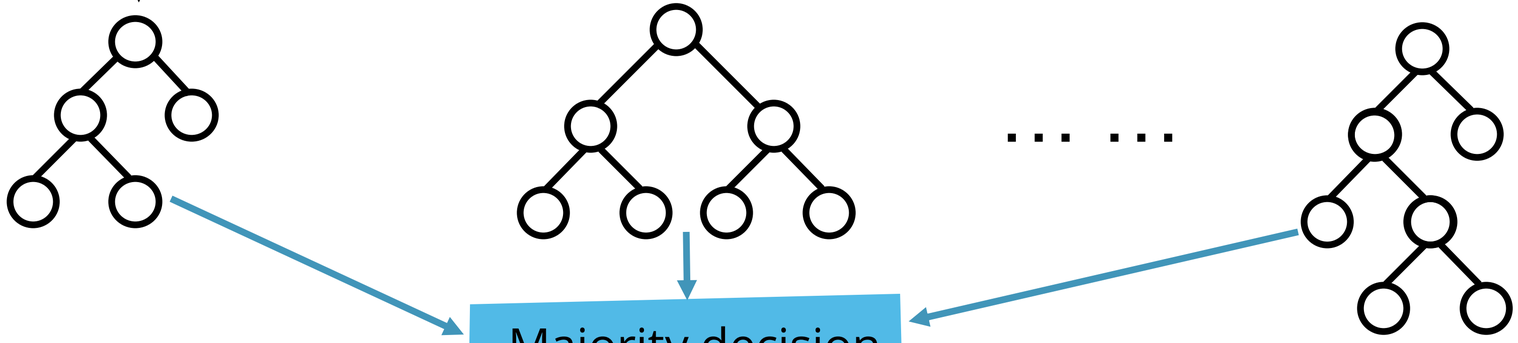
Aggregation

X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2



Majority decision

TIC em trilhas

Random Forests - Funcionamento

- **Passo 1**: criação do bootstrap dataset
  - Considere o seguinte dataset:

Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Sim	Não	Sim	125	Sim
Não	Sim	Não	180	Não
Não	Não	Sim	210	Não
Sim	Não	Sim	130	Sim

Random Forests - Funcionamento

- **Passo 1**: criação do bootstrap dataset
  - Geração de diferentes subsets de forma aleatória a partir do dataset original.

Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Sim	Não	Sim	125	Sim
Não	Sim	Não	180	Não
Não	Não	Sim	210	Não
Sim	Não	Sim	130	Sim

Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	125	Sim

Bootstrap Dataset (Bagging)

Random Forests - Funcionamento

- **Passo 2**: criação das árvores de decisão
  - A partir de cada subset seleciona um número X de atributos aleatoriamente.

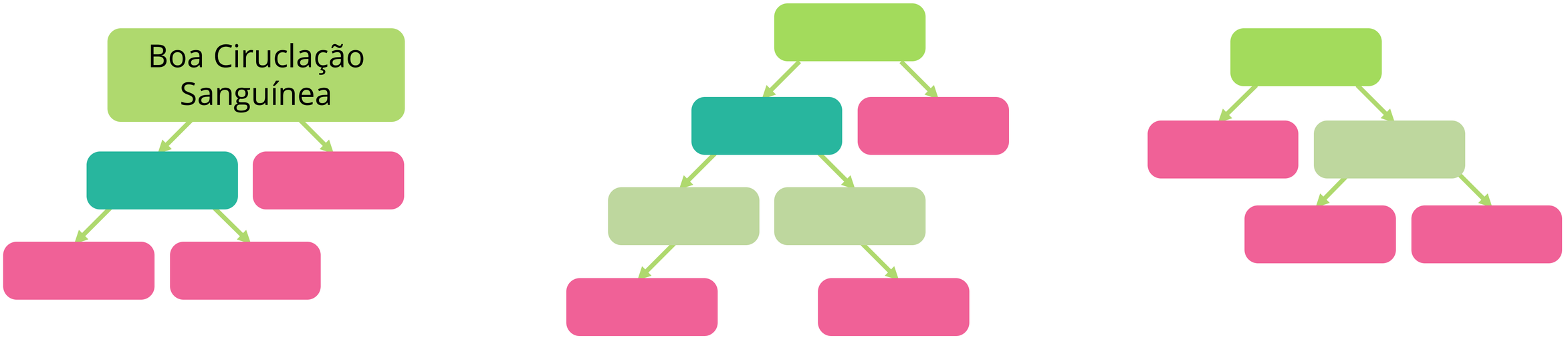
Boa circulação sanguínea	Arterias bloqueadas
Sim	Não
Não	Sim
Não	Sim

Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim

Bootstrap Dataset (Bagging)

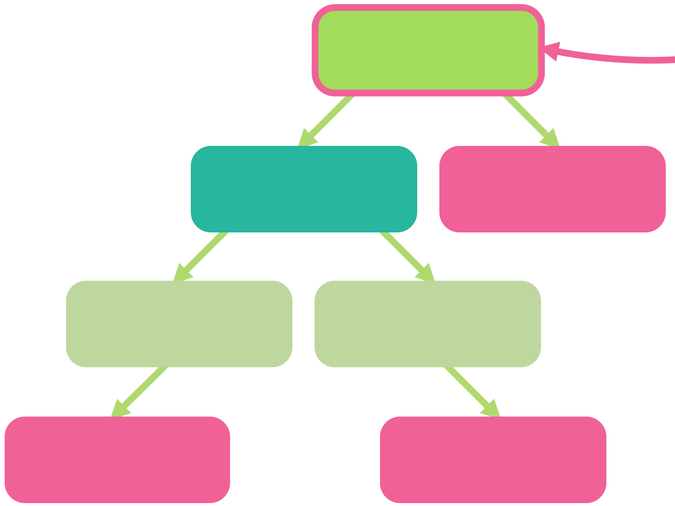
Random Forests - Funcionamento

- **Passo 2**: criação das árvores de decisão
  - As árvores são construídas considerando apenas os subconjuntos de atributos selecionados.
  - Cada árvore tem um tamanho distinto pois foi escolhido um conjunto de atributos diferentes, sendo esse o objetivo da “floresta aleatória”.
  - Gera modelos distintos e a combinação deles vai tornar um modelo mais robusto e assertivo



Random Forests - Funcionamento

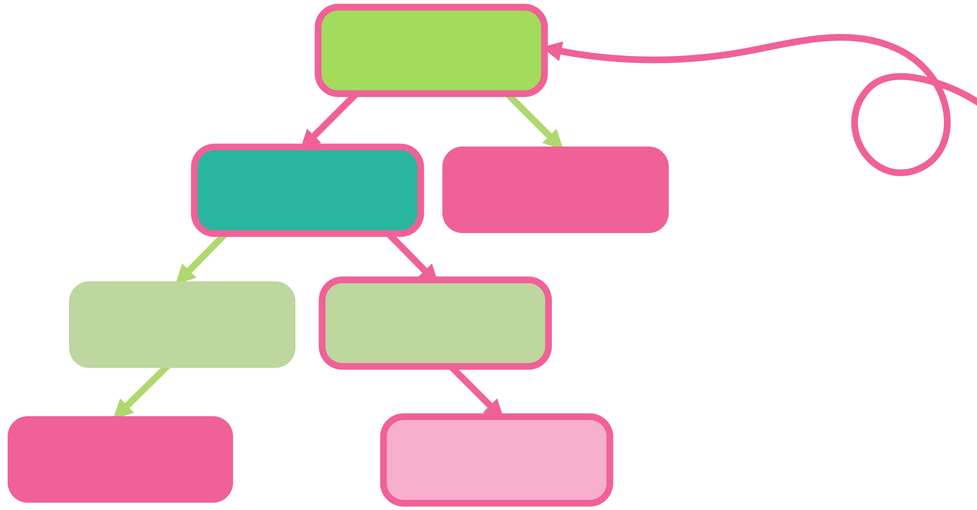
- **Passo 3**: classificação dos dados de teste/predição
  - Consulta todas as árvores da floresta para classificar (Sim/Não)
  - Comparação com cada nó de cada árvore da floresta.



Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	

Random Forests - Funcionamento

- **Passo 3**: classificação dos dados de teste/predição
  - Árvore 1: Percorre toda a árvore até o seu nodo folha para descobrir a classe.



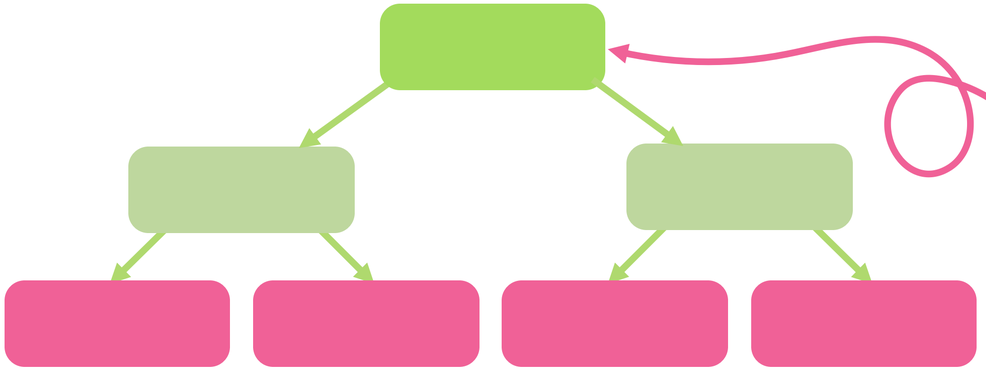
Não

Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	

Doença cardiaca	
Sim	Não
0	1

Random Forests - Funcionamento

- **Passo 3**: classificação dos dados de teste/predição
  - Árvore 2: Começa pelo atributo que está no nodo raiz.

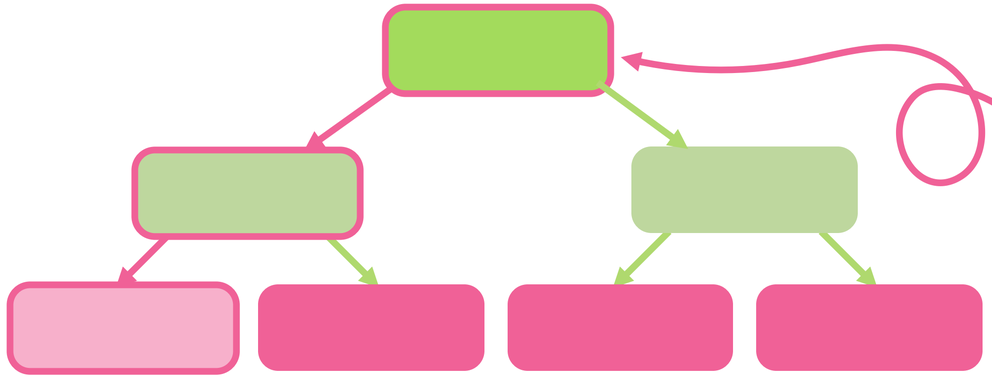


Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	

Doença cardiaca	
Sim	Não
0	1

Random Forests - Funcionamento

- **Passo 3**: classificação dos dados de teste/predição
  - Árvore 2: Percorre toda a árvore até o seu nodo folha para descobrir a classe.



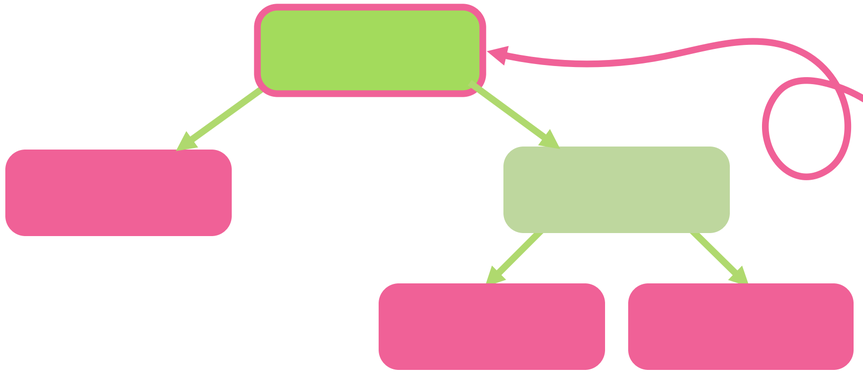
Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	

Sim

Doença cardiaca	
Sim	Não
1	1

Random Forests - Funcionamento

- **Passo 3**: classificação dos dados de teste/predição
  - Árvore 3: Começa pelo atributo que está no nodo raiz.

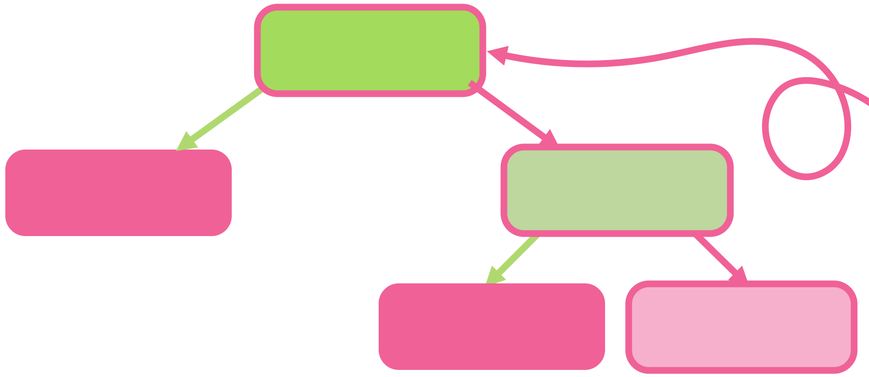


Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	

Doença cardiaca	
Sim	Não
1	1

Random Forests - Funcionamento

- **Passo 3**: classificação dos dados de teste/predição
  - Classe doença cardíaca mais votada: Não
  - Votação majoritária (*bagging*)

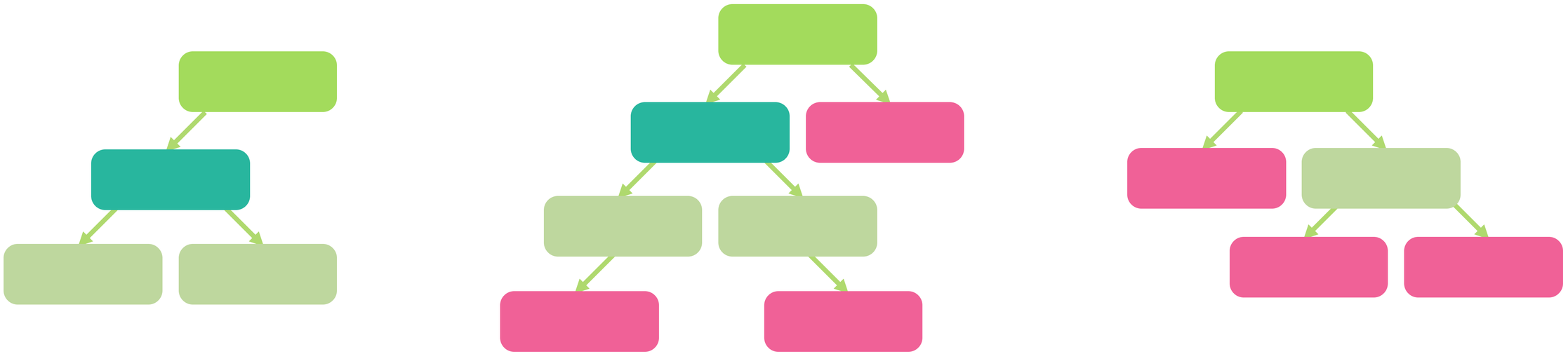


Dor no peito	Boa circulação sanguínea	Arterias bloqueadas	Peso	Doença cardiaca
Não	Sim	Não	180	

Doença cardiaca	
Sim	Não
1	2

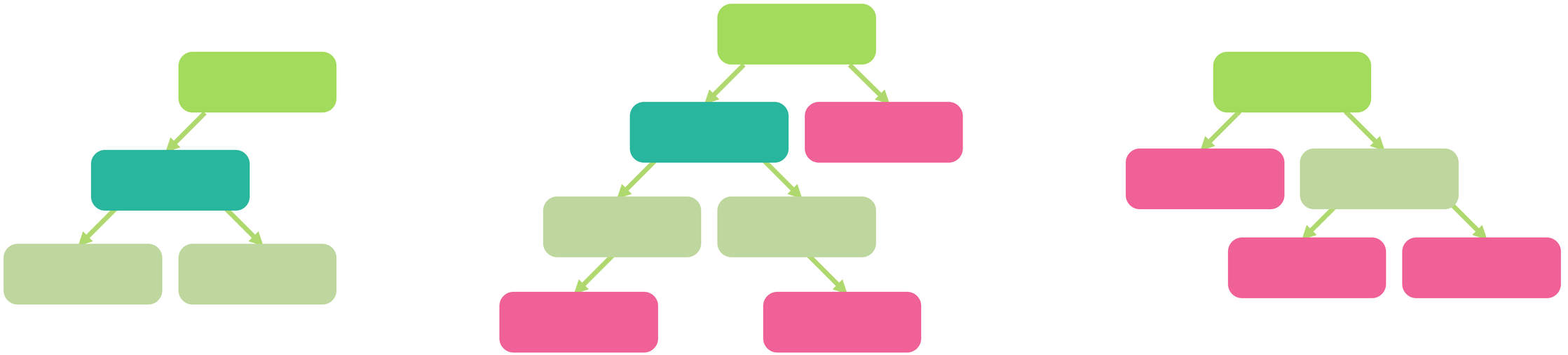
Random Forests

- **Vantagem**
  - Maior robustez
  - Menos propenso a sofrer *overfitting* em comparação com uma única árvore de decisão



Random Forests

- **Desvantagem**
  - Exige um maior poder de processamento devido a sua robustez
  - O processo de classificação de novas amostras pode ser lento (no caso quando estiver em produção).



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.