

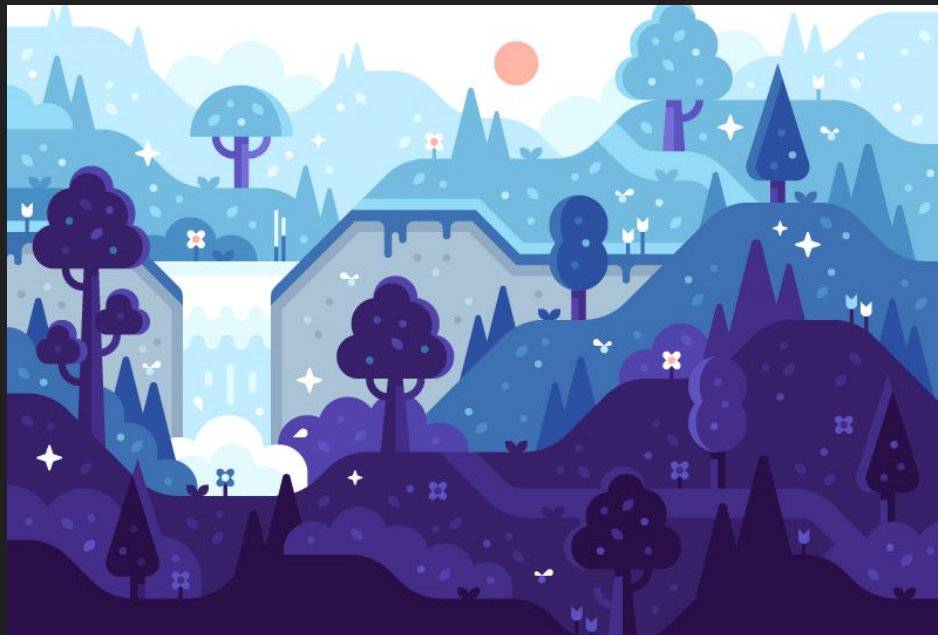
Data Science do ZERO

Capítulo 06 - Machine Learning

Random Forest
(Floresta Aleatória)

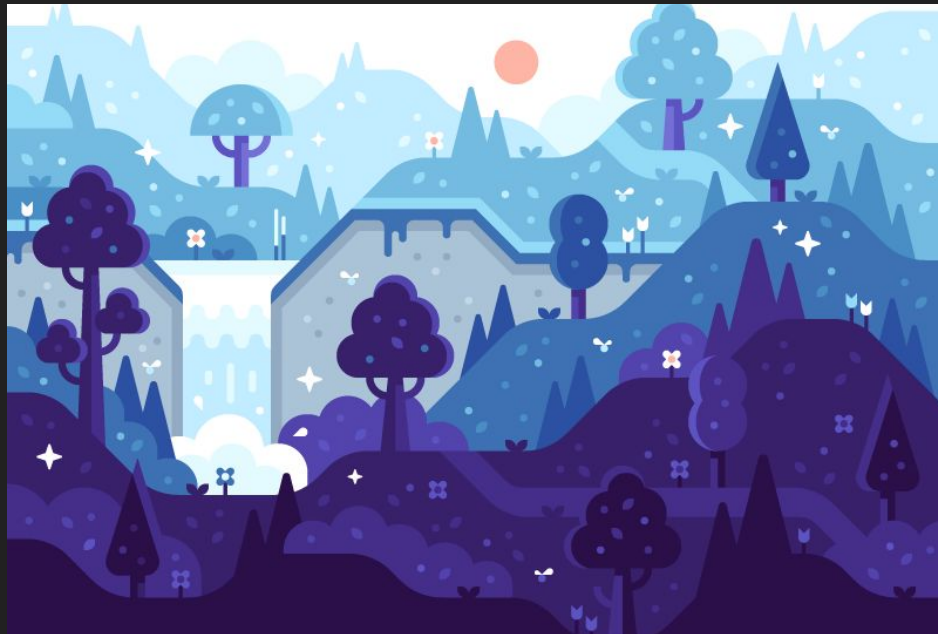
Random Forest

- Algoritmo de **Machine Learning** Supervisionado utilizado para Classificação ou regressão.
- Combinação da simplicidade das árvores de decisão com a flexibilidade e aleatoriedade para melhorar a precisão



Random Forest

- Dezenas de árvores combinadas para predizer o melhor resultado.
- Aleatoriedade na seleção de atributos ao invés da seleção a partir do cálculo de impureza.



Random Forest

- Primeiro passo, criação do **bootstrap dataset**.

Dor no peito	Boa Circulação Sanguínea	Arterias Bloqueadas	Peso	Doença Cardíaca
Sim	Não	Sim	125	Sim
Não	Sim	Não	180	Não
Não	Não	Sim	210	Não
Sim	Não	Sim	130	Sim

Random Forest

Dor no peito	Boa Circulação Sanguínea	Arterias Bloqueadas	Peso	Doença Cardíaca
Sim	Não	Sim	125	Sim
Não	Sim	Não	180	Não
Não	Não	Sim	210	Não
Sim	Não	Sim	130	Sim

Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardíaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim



Bootstrap Dataset

Random Forest

A partir do conjunto original ..
Selecione um número N de
features aleatoriamente

Boa circ Sanguínea	Arterias Bloq.
Sim	Não
Não	Sim
Não	Sim

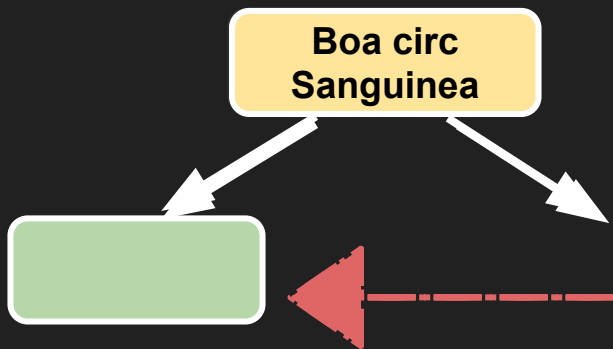
Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardíaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim



Bootstrap Dataset

Random Forest

A partir do subconjunto selecionado é feita a verificação do atributo que melhor separa os dados..



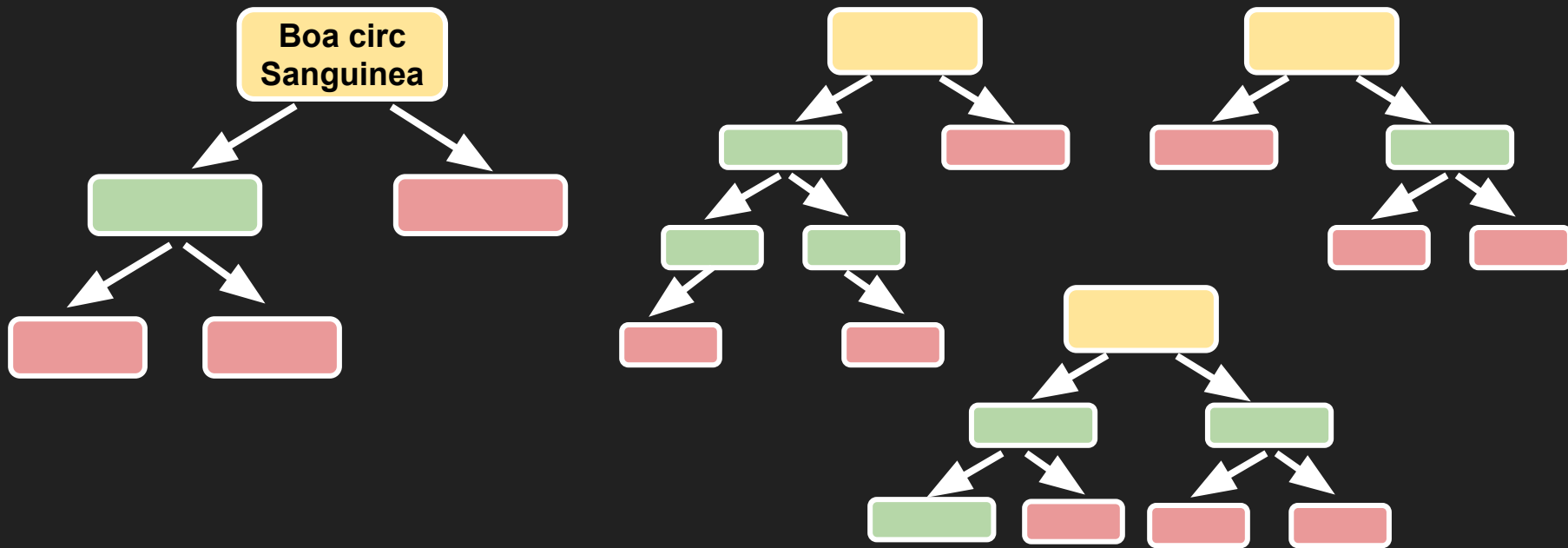
Dor no peito	Boa circ Sanguinea	Arterias Bloq.	Peso	Doença Cardíaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim

Agora é preciso separar mais 2 atributos a partir dos três resultantes para separar os dados e construir a árvore.

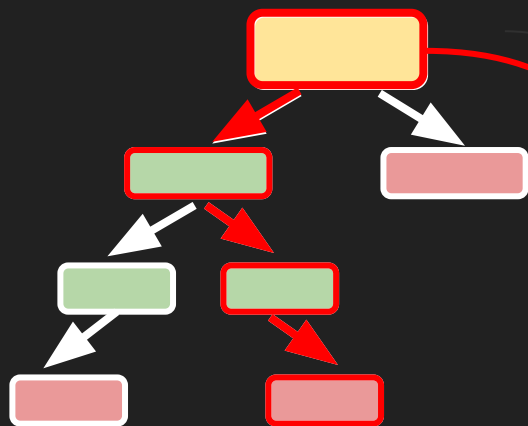

Bootstrap Dataset

Random Forest

As árvores são construídas considerando apenas os **subconjuntos de atributos** selecionados.



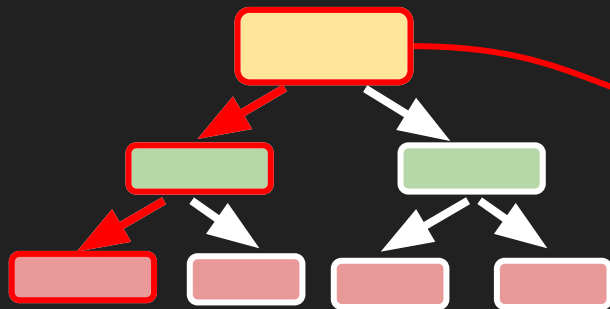
Random Forest



Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardíaca
Não	Sim	Não	180	

Doença Cardíaca	
SIM	NÃO
0	1

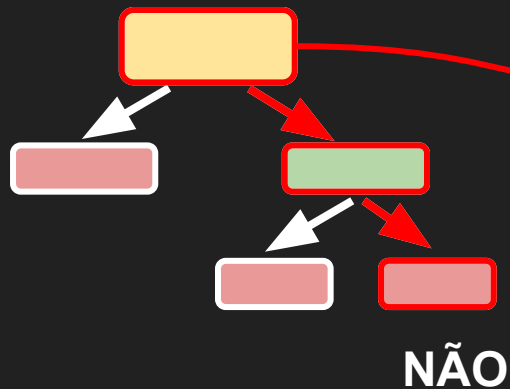
Random Forest



Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardíaca
Não	Sim	Não	180	

Doença Cardíaca	
SIM	NÃO
0	1

Random Forest



Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardíaca
Não	Sim	Não	180	

Doença Cardíaca	
SIM	NÃO
1	2

Random Forest

1. Criação do Bootstrapped Dataset
2. A cada passo é selecionado um número N de features para montar a árvore.
3. Diversas árvores são criadas a partir de subconjuntos diferentes.
4. A instância de teste deve percorrer cada árvore da floresta e a classe definida será a mais votada.



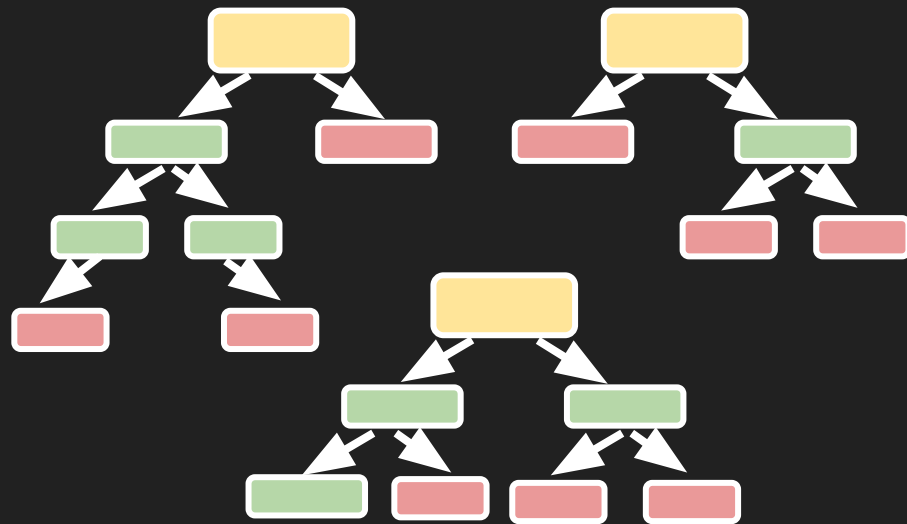
Random Forest

- Algumas vantagens
 - Maior robustez
 - Menos propenso a sofrer Overfitting em comparação com uma única Árvore de Decisão
 - Permite a descoberta de conhecimento.
 - Poucos parâmetros para ajustes.



Random Forest

- Desvantagens
 - Exige um maior poder de processamento.
 - Pode ser lento o processo de classificação de novas amostras.



Hands on!