
Aprendizado de Máquina II

Radom Forest



Prof^a. Renata De Paris

Especialização em Ciência de Dados

Roteiro da Aula

- ❑ Definição
- ❑ Representação
- ❑ Funcionamento
- ❑ Aplicações – casos de uso
- ❑ Atividade

Random Forests

■ Definição

- ❑ Melhora o desempenho de generalização construindo um conjunto de árvores de decisão descorelacionadas.
- ❑ Utiliza uma amostra de bootstrap diferente de treinamento de dados para aprender a partir de árvores de decisões.
- ❑ Constrói conjuntos (ensembles) de árvores de decisão manipulando instâncias e atributos de entrada.
- ❑ Para cada nodo, o melhor critério de divisão é escolhido entre um pequeno conjunto de atributos selecionados randomicamente.
- ❑ Combina simplicidade das árvores de decisão com a flexibilidade e aleatoriedade para melhorar a precisão
- ❑ É um tipo de *ensemble learning* usado especificamente para árvores de decisão.

Random Forests - Representação

Training dataset

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

Bootstrap

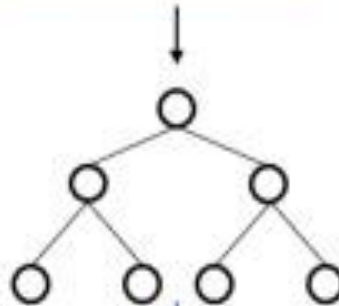
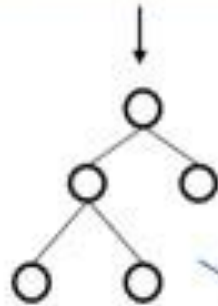
X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

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

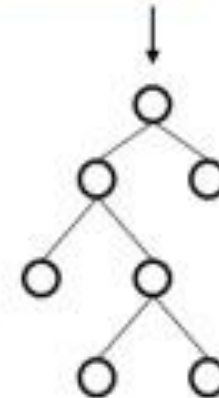
...

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

Ensemble of trees



...



Aggregation

Majority decision

Random Forests - Funcionamento

- **Passo 1:** criação do bootstrap dataset

- Considere o seguinte 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

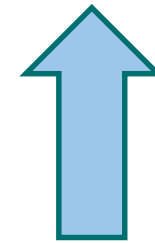
Random Forests - Funcionamento

■ 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 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 Circulação Sanguínea	Arterias Bloqueadas	Peso	Doença Cardíaca
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

- A partir de cada subset seleciona um número X de atributos aleatoriamente.

Boa Circulação Sanguínea	Arterias Bloqueadas
Sim	Não
Não	Sim
Não	Sim

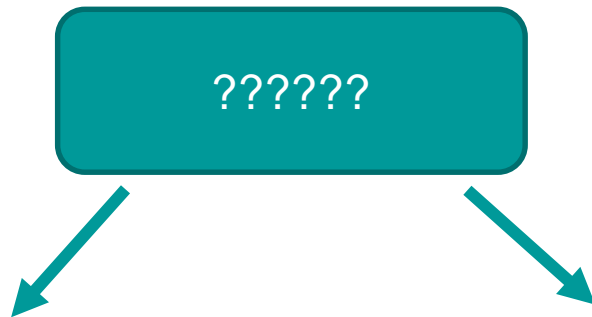
Dor no peito	Boa Circulação Sanguínea	Arterias Bloqueadas	Peso	Doença Cardíaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim

↑
Bootstrap Dataset
(Bagging)

Random Forests - Funcionamento

■ Passo 2: criação das árvores de decisão

- A partir do subset selecionado é feita a verificação do atributo que melhor separa os dados por meio de métricas para avaliar o grau de impureza.



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

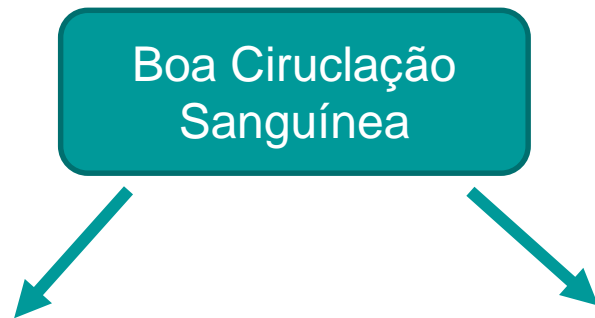
Bootstrap Dataset
(Bagging)

A large blue arrow with a teal outline points upwards from the text "Bootstrap Dataset (Bagging)" to the table above it.

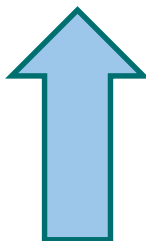
Random Forests - Funcionamento

■ Passo 2: criação das árvores de decisão

- A partir do subset selecionado é feita a verificação do atributo que melhor separa os dados por meio de métricas para avaliar o grau de impureza.



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

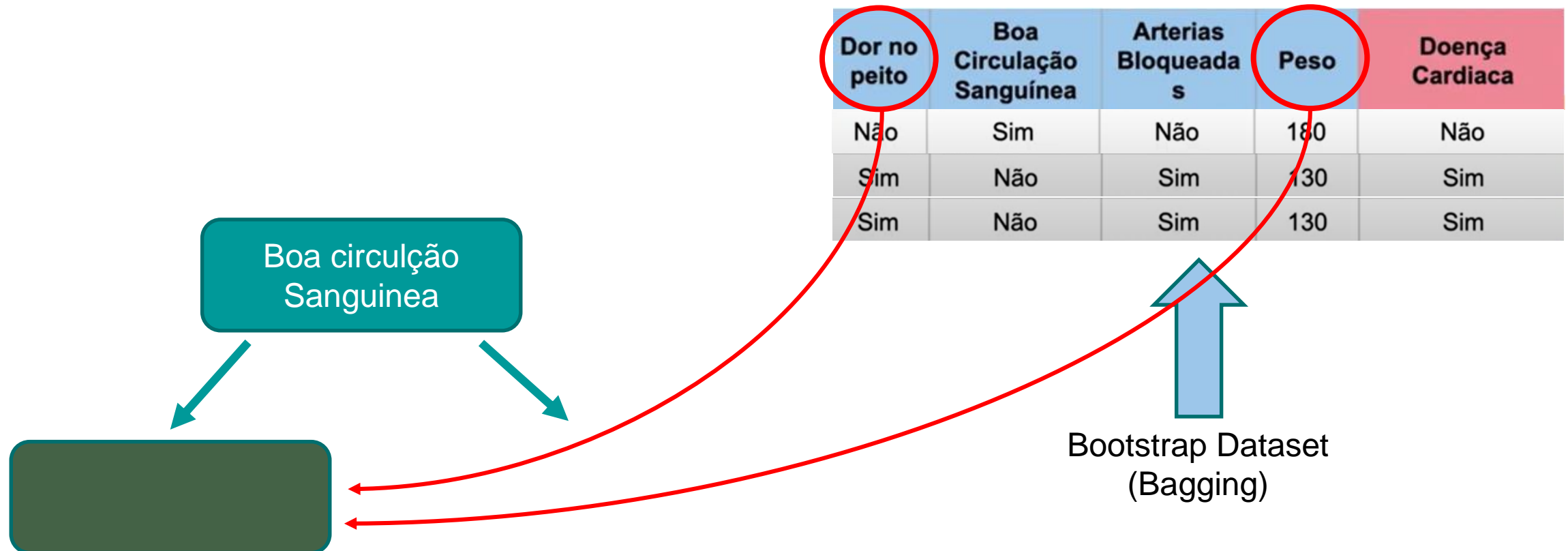


Bootstrap Dataset
(Bagging)

Random Forests - Funcionamento

■ Passo 2: criação das árvores de decisão

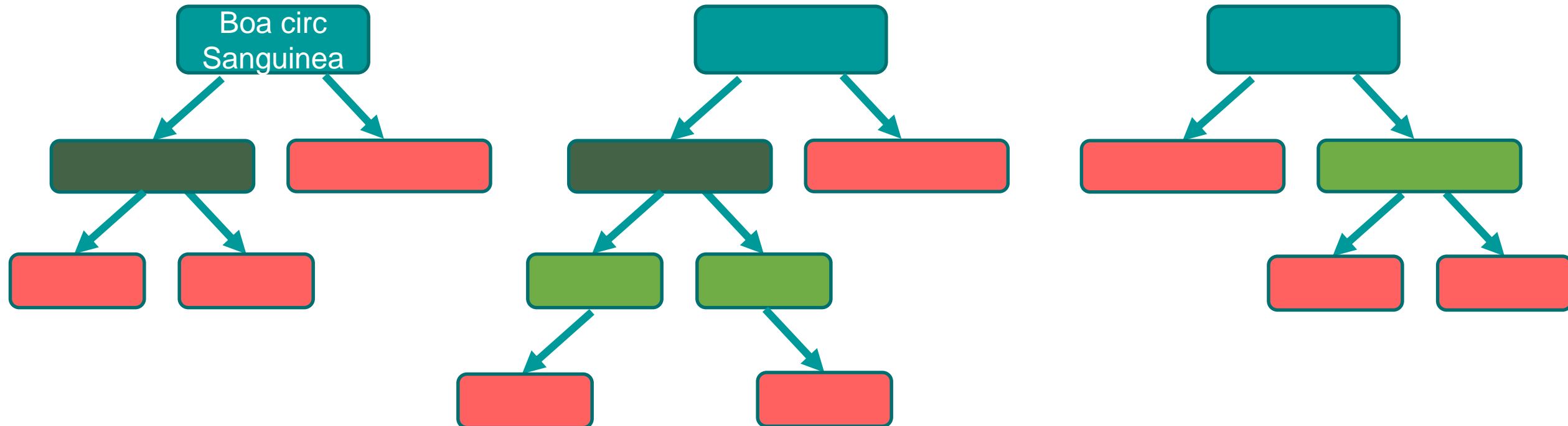
- A partir dos três atributos restante são escolhidos mais 2 aleatoriamente para separar os dados e construir a árvore.



Random Forests - Funcionamento

■ Passo 2: criação das árvores de decisão

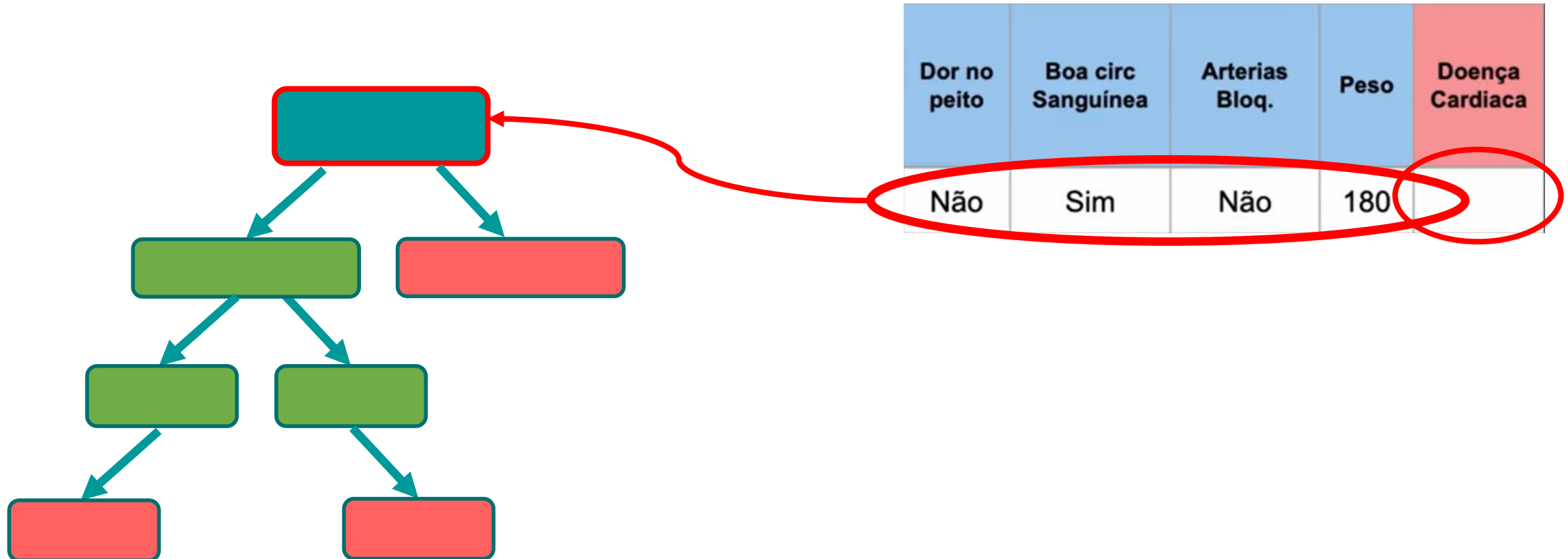
- As árvores são construídas considerando apenas os subconjuntos de atributos selecionados.
- Cada árvore tem um tamanho distinto pois foi escolhido um conjunto de atributos diferentes, sendo esse o objetivo da “floresta aleatória”.
- Gera modelos distintos e a combinação deles vai tornar um modelo mais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



Random Forests - Funcionamento

■ Passo 3: classificação dos dados de teste/predição

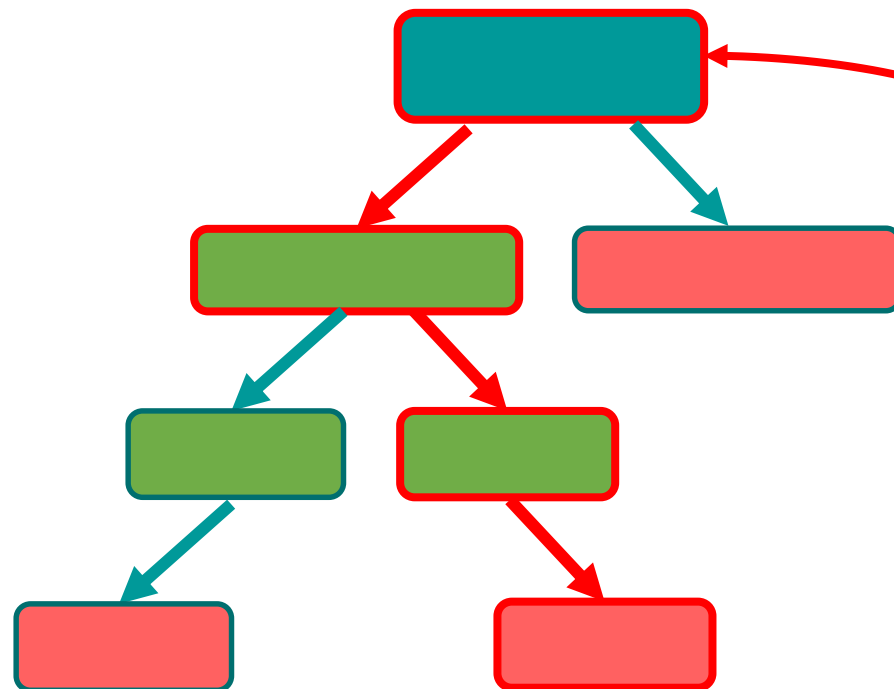
- ❑ Consulta todas as árvores da floresta para classificar (Sim/Não)
- ❑ Comparação com cada nó de cada árvore da floresta.



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.



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

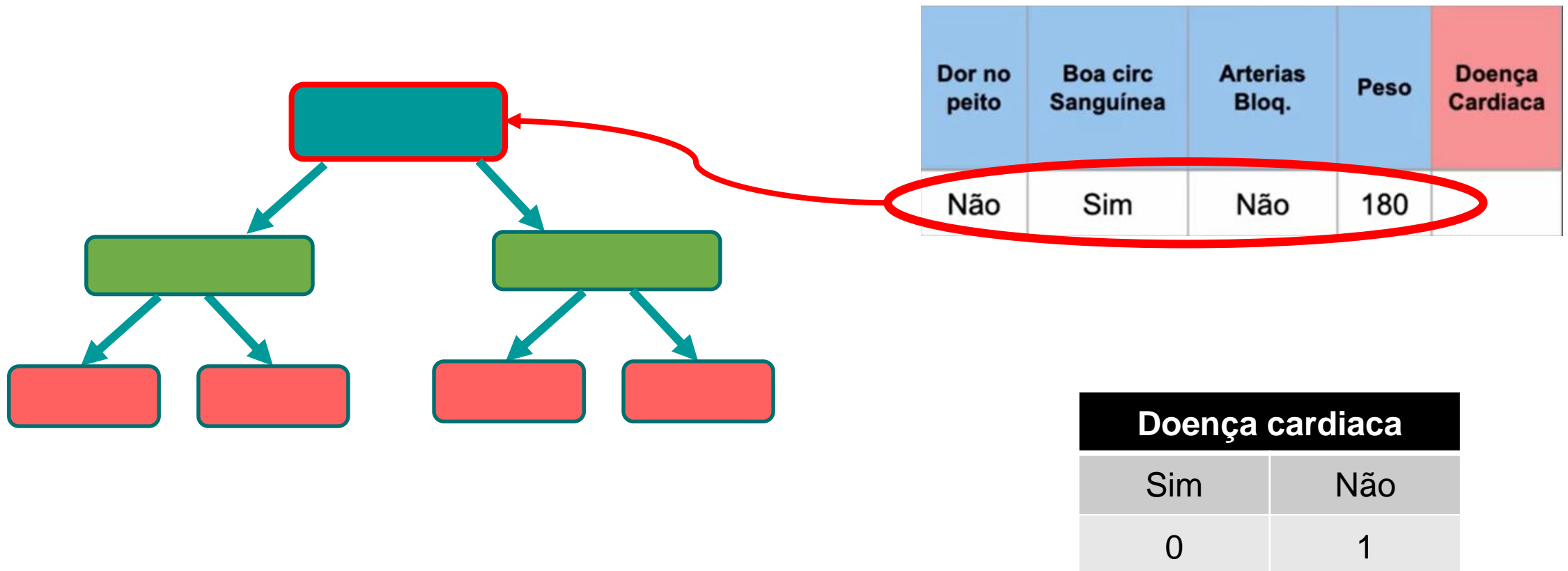
Doença cardíaca	
Sim	Não
0	1

Não

Random Forests - Funcionamento

■ **Passo 3:** classificação dos dados de teste/predição

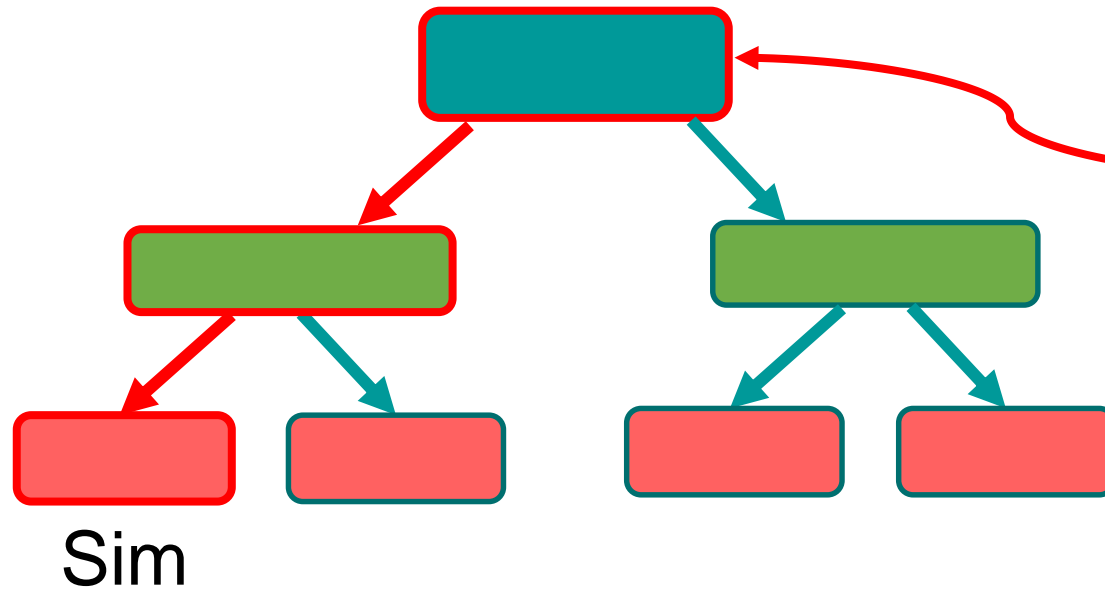
- **Árvore 2:** Começa pelo atributo que está no nodo raiz.



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.

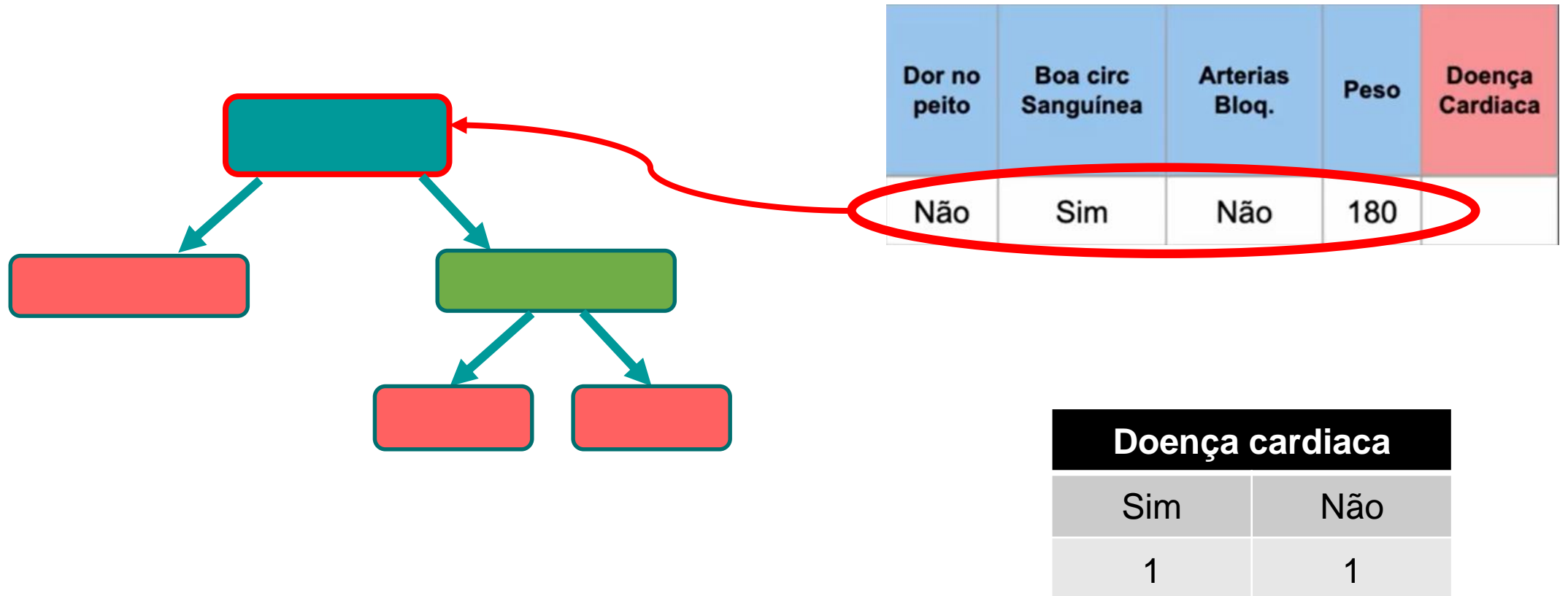


Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardíaca
Não	Sim	Não	180	

Doença cardíaca	
Sim	Não
1	1

Random Forests - Funcionamento

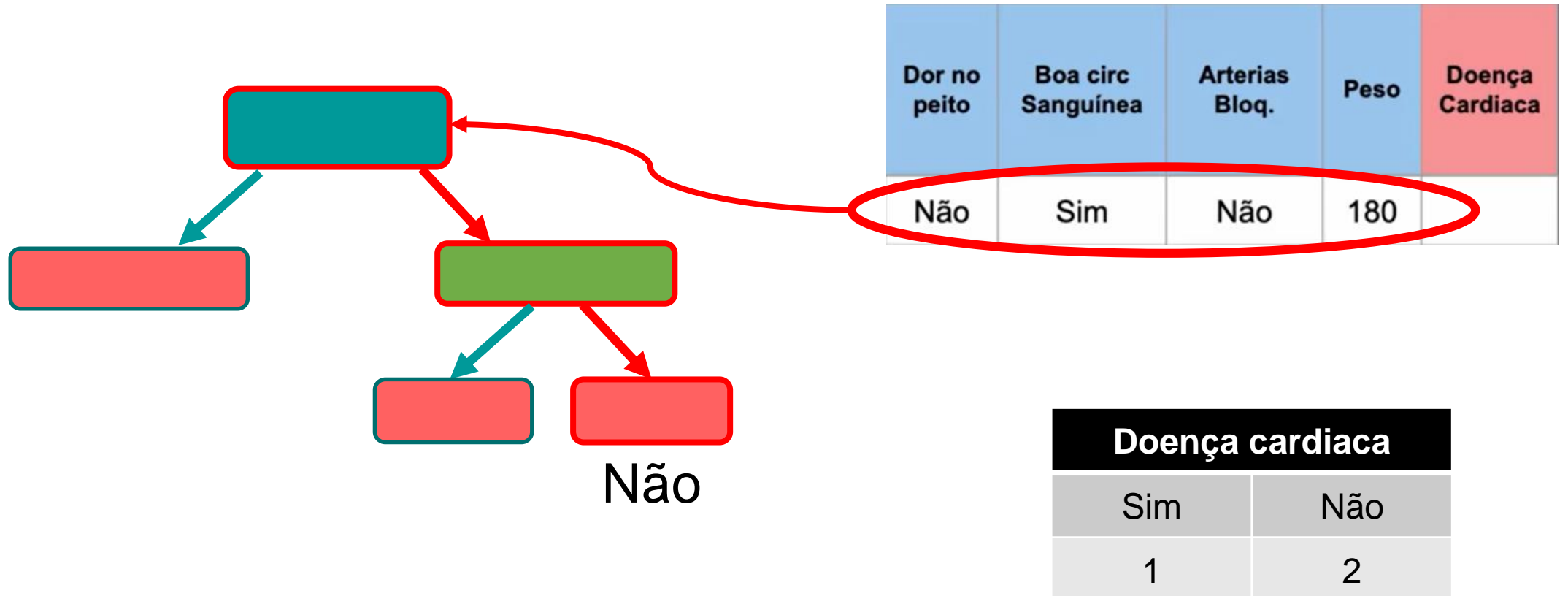
- **Passo 3:** classificação dos dados de teste/predição
 - Árvore 3: Começa pelo atributo que está no nodo raiz.



Random Forests - Funcionamento

■ Passo 3: classificação dos dados de teste/predição

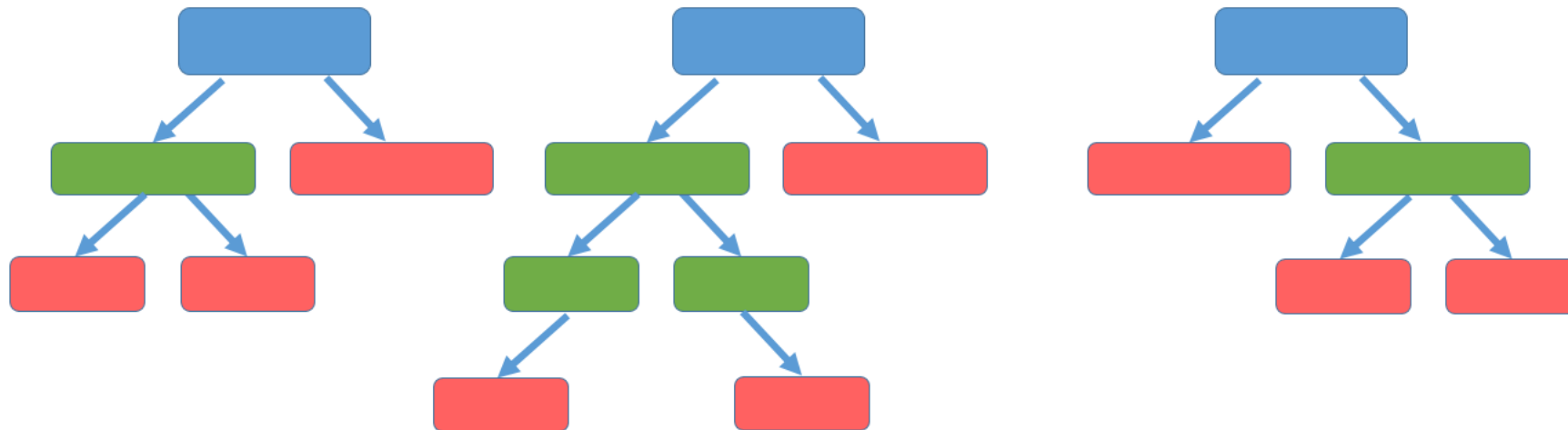
- ❑ Classe doença cardíaca mais votada: Não
- ❑ Votação majoritária (*bagging*)



Random Forests

■ Vantagem

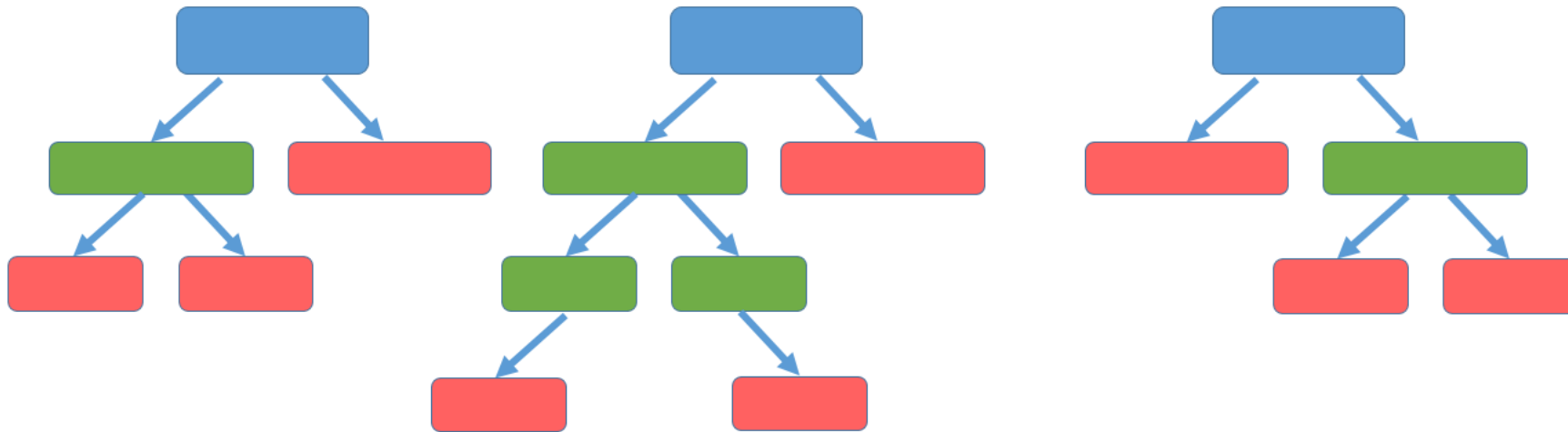
- ❑ Maior robustez
- ❑ Menos propenso a sofrer *overfitting* em comparação com uma única árvore de decisão
- ❑ Permite a descoberta de conhecimento
- ❑ Poucos parâmetros para ajustes



Random 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).



Random Forest - Exemplo de Uso

Chaos, Solitons and Fractals 140 (2020) 110210



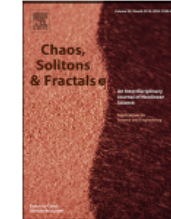
ELSEVIER

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

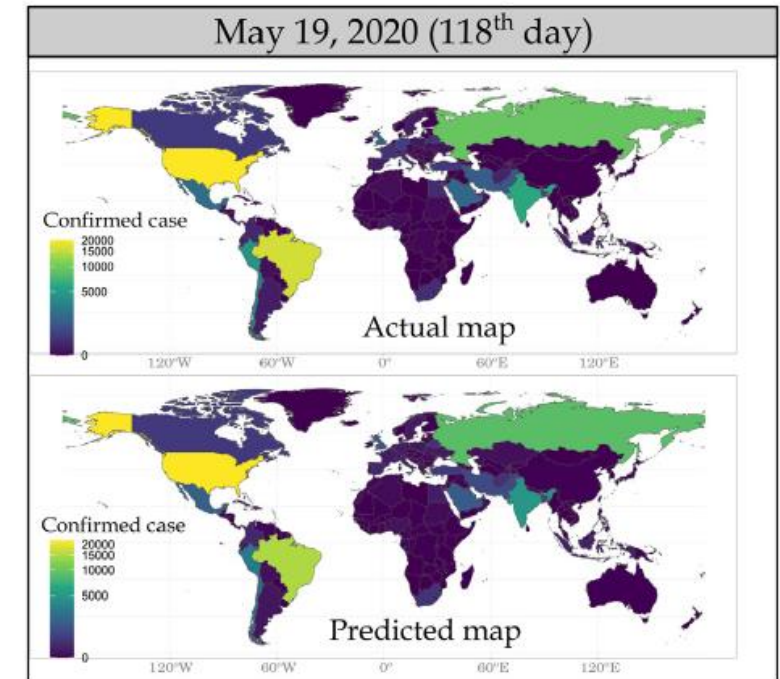
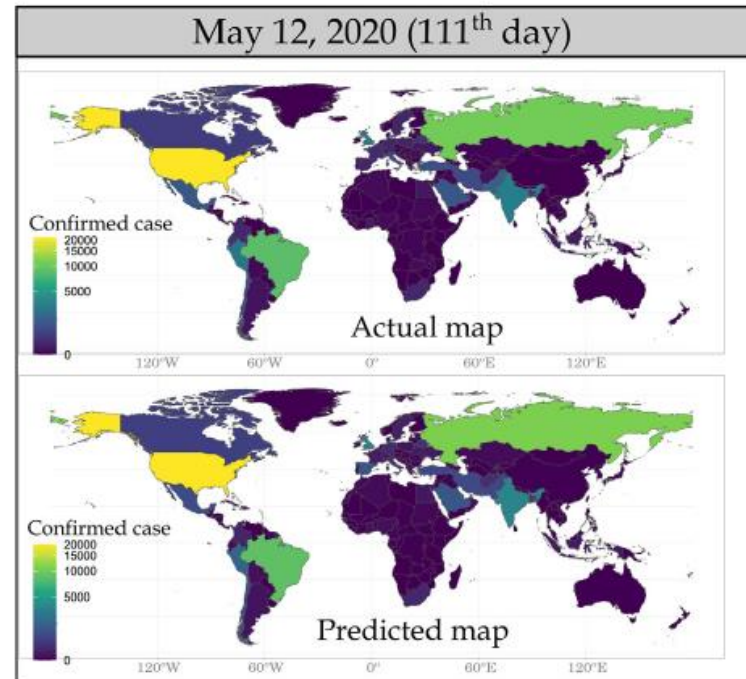


Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm



Cafer Mert Yeşilkanat

Science Teaching Department, Artvin Çoruh University, Artvin, Turkey



Random Forest - Exemplo de Uso

Yao et al. *BMC Bioinformatics* (2020) 21:126
<https://doi.org/10.1186/s12859-020-3458-1>


BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

A random forest based computational model for predicting novel lncRNA-disease associations



Dengju Yao^{1*} , Xiaojuan Zhan², Xiaorong Zhan³, Chee Keong Kwoh⁴, Peng Li¹ and Jinke Wang⁵

* Correspondence: ydkvictory@hrbust.edu.cn

¹School of Software and Microelectronics, Harbin University of Science and Technology, Harbin 150080, China

Full list of author information is available at the end of the article

Abstract

Background: Accumulated evidence shows that the abnormal regulation of long non-coding RNA (lncRNA) is associated with various human diseases. Accurately identifying disease-associated lncRNAs is helpful to study the mechanism of lncRNAs in diseases and explore new therapies of diseases. Many lncRNA-disease association (LDA) prediction models have been implemented by integrating multiple kinds of data resources. However, most of the existing models ignore the interference of noisy and redundancy information among these data resources.

Random Forest - Exemplo de Uso

Random Forest for Credit Card Fraud Detection

Shiyang Xuan

*Department of Computer Science
Tongji University
Shanghai, China
xsyfor@tongji.edu.cn*

Guanjun Liu

*Department of Computer Science
Tongji University
Shanghai, China
liuguanjun@tongji.edu.cn*

Zhenchuan Li

*Department of Computer Science
Tongji University
Shanghai, China
1510482@tongji.edu.cn*

Lutao Zheng

*Department of Computer Science
Tongji University
Shanghai, China
zhenglutao103@163.com*

Shuo Wang

*Department of Computer Science
Tongji University
Shanghai, China
wangshuo@tongji.edu.cn*

Changjun Jiang

*Department of Computer Science
Tongji University
Shanghai, China
cjjiang@tongji.edu.cn*

TABLE III
RESULTS OF TWO KINDS RANDOM FORESTS

Measure Models	Accuracy	Precision	Recall	F-Measure
Random Forest I	91.96%	90.27%	67.89%	0.7811
Random Forest II	96.77%	89.46%	95.27%	0.9601

Random Forest - Exemplo de Uso

Received: 28 July 2016 | Revised: 25 April 2017 | Accepted: 8 May 2017

DOI: 10.1002/sam.11348

ORIGINAL ARTICLE

WILEY

Random forest missing data algorithms

Fei Tang | Hemant Ishwaran

Division of Biostatistics, University of Miami
Coral Gables, Florida

Correspondence

Hemant Ishwaran, Division of Biostatistics,
University of Miami, 1320 S Dixie Hwy, Coral
Gables, FL 33146. Email:
hemant.ishwaran@gmail.com

Funding Information

National Institutes of Health
R01CA163739

Random forest (RF) missing data algorithms are an attractive approach for imputing missing data. They have the desirable properties of being able to handle mixed types of missing data, they are adaptive to interactions and nonlinearity, and they have the potential to scale to big data settings. Currently there are many different RF imputation algorithms, but relatively little guidance about their efficacy. Using a large, diverse collection of data sets, imputation performance of various RF algorithms was assessed under different missing data mechanisms. Algorithms included proximity imputation, on the fly imputation, and imputation utilizing multivariate unsupervised and supervised splitting—the latter class representing a generalization of a new promising imputation algorithm called missForest. Our findings reveal RF imputation to be generally robust with performance improving with increasing correlation. Performance was good under moderate to high missingness, and even (in certain cases) when data was missing not at random.

Atividade para entregar

- A partir do **dataset escolhido para o trabalho final da disciplina** e utilizando a biblioteca scikitlearn da linguagem de programação Python, realize as seguintes tarefas:
 1. Execute o algoritmo Random Forest do sklearn com as seguintes configurações:
 - 1.1 Aplique o método cross-validation para 10 k-folds (cv =10)
 - 1.2 Altere o parâmetro quantidade de árvores geradas na floresta (n_estimator), podendo ser de 100 a 1000.
 - 1.3 Encontre a melhor configuração utilizando as métricas Acuracy, Precision e Recall.
 2. Execute novamente o algoritmo Random Forest com o método holdout utilizando entre 25 a 30% dos dados para teste, utilizando como parâmetro do algoritmo o n_estimator que apresentou o melhor desempenho no exercício 1.
 - 2.1 Analise e compare os resultados obtidos, utilizando a matriz de confusão para interpretar os valores de *Acuracy*, *Precision*, *Recall* e *F1-score* do dataset de teste.
 - 2.2 O objetivo é encontrar a melhor parametrização do algoritmo Random Forest para o dataset utilizado, baseando-se nas medidas de avaliação de métodos supervisionados.

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.