

 [View PDF](#)

[Download full issue](#)

Outline

[Highlights](#)

[Abstract](#)

[Keywords](#)

[1. Introduction](#)

[2. Methods](#)

[3. Results](#)

[4. Discussion](#)

[5. Conclusions](#)

[Declaration of Competing Interest](#)

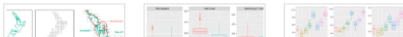
[Appendix A. Supplementary data](#)

[References](#)

[Show full outline](#) 

[Cited by \(18\)](#)

[Figures \(6\)](#)



Ecosystem Services

Volume 46, December 2020, 101212



Promotion of afforestation in New Zealand's marginal agricultural lands through payments for environmental services

Thales A.P. West ^{a, b} , Juan J. Monge ^c, Les J. Dowling ^d, Steve J. Wakelin ^a, Holly K. Gibbs ^{e, f}

[Show more](#) 

[+ Add to Mendeley](#) [Share](#) [Cite](#)

<https://doi.org/10.1016/j.ecoser.2020.101212> 

[Get rights and content](#) 

 Full text access

Highlights

- 56% of marginal agricultural lands present lower economic returns than forestry.
- 77–100% of marginal agricultural lands could financially benefit

Recommended articles

[Are forest landowners ready for woody energy crops? Preferences for afforestation...](#)

Energy Economics, Volume 73, 2018, pp. 239-247

Mario Soliño, ..., Alejandro Caparrós

 [View PDF](#)

[Impact on profitability, risk, optimum rotation age and afforestation of changing...](#)

Forest Policy and Economics, Volume 116, 2020, Article...

Bruce Manley

 [View PDF](#)

[Determining the balance between active and passive indigenous forest restoration...](#)

Forest Ecology and Management, Volume 479, 2021, A...

Adam S. Forbes, ..., Lance Taurua

 [View PDF](#)

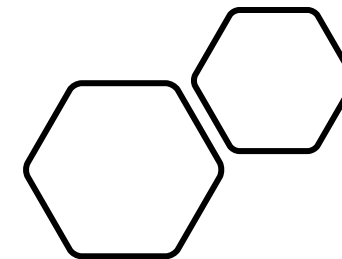
[Show 3 more articles](#) 

Article Metrics

Citations

 [View PDF](#)

 [FEEDBACK](#)



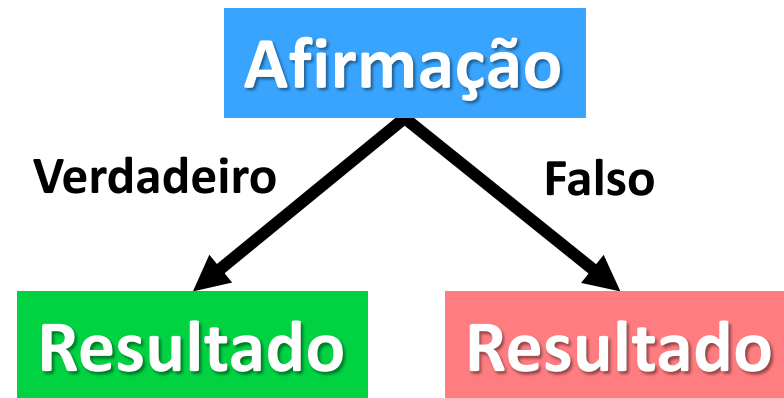
Introdução ao
aprendizado de
máquina:

Florestas Aleatórias (FA)

Random Forests (RF)

Thales A. P. West, Ph.D.

Árvores de decisão





What Causes Wildfires?

A spark, in the presence of fuel and oxygen, can cause a fire, which can further spread depending on various weather conditions.

A SPARK

Fires can be sparked by natural causes, but most wildfires in the U.S. are caused by human activity.

Natural causes:



Lightning strikes



Heat from the sun

Human activity:



Vehicles & machinery



Cigarette butts



Fireworks



Campfires & burning brush

+

FUEL & OXYGEN

Fuel is usually dry vegetation:



Grass



Brush



Trees

Oxygen is almost everywhere.



=

FIRE!

X

WEATHER CONDITIONS

Certain weather conditions can make fires bigger, faster, and more dangerous.



Strong wind



Drought

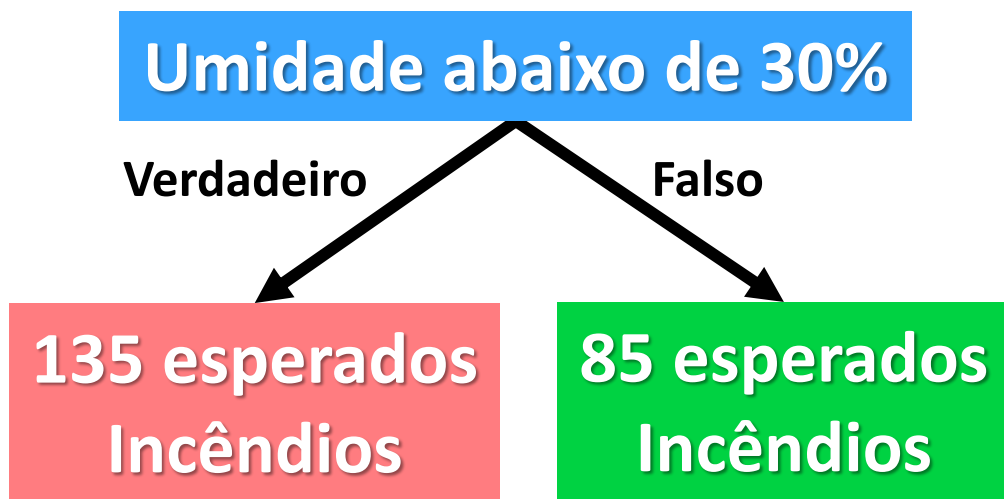


High temperatures & low humidity

weather.gov

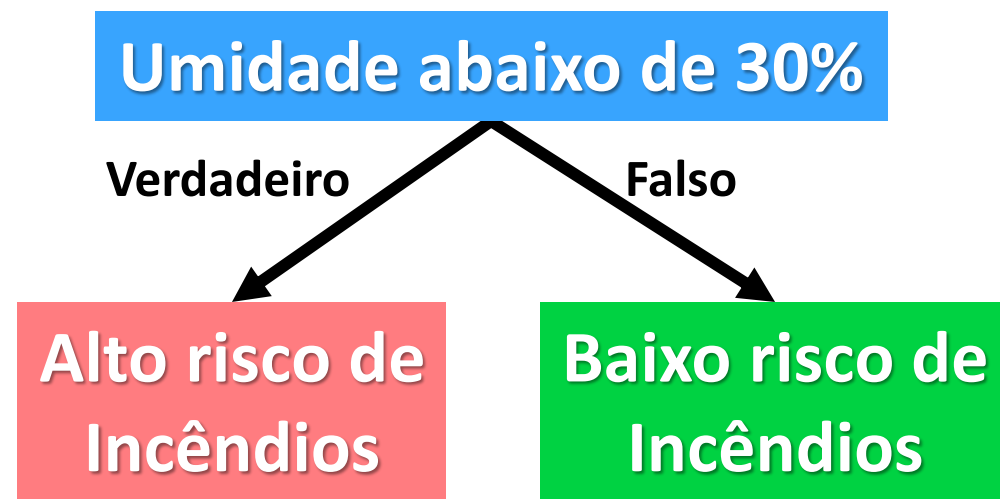


Árvores de decisão



Esta árvore prevê valores numéricos...

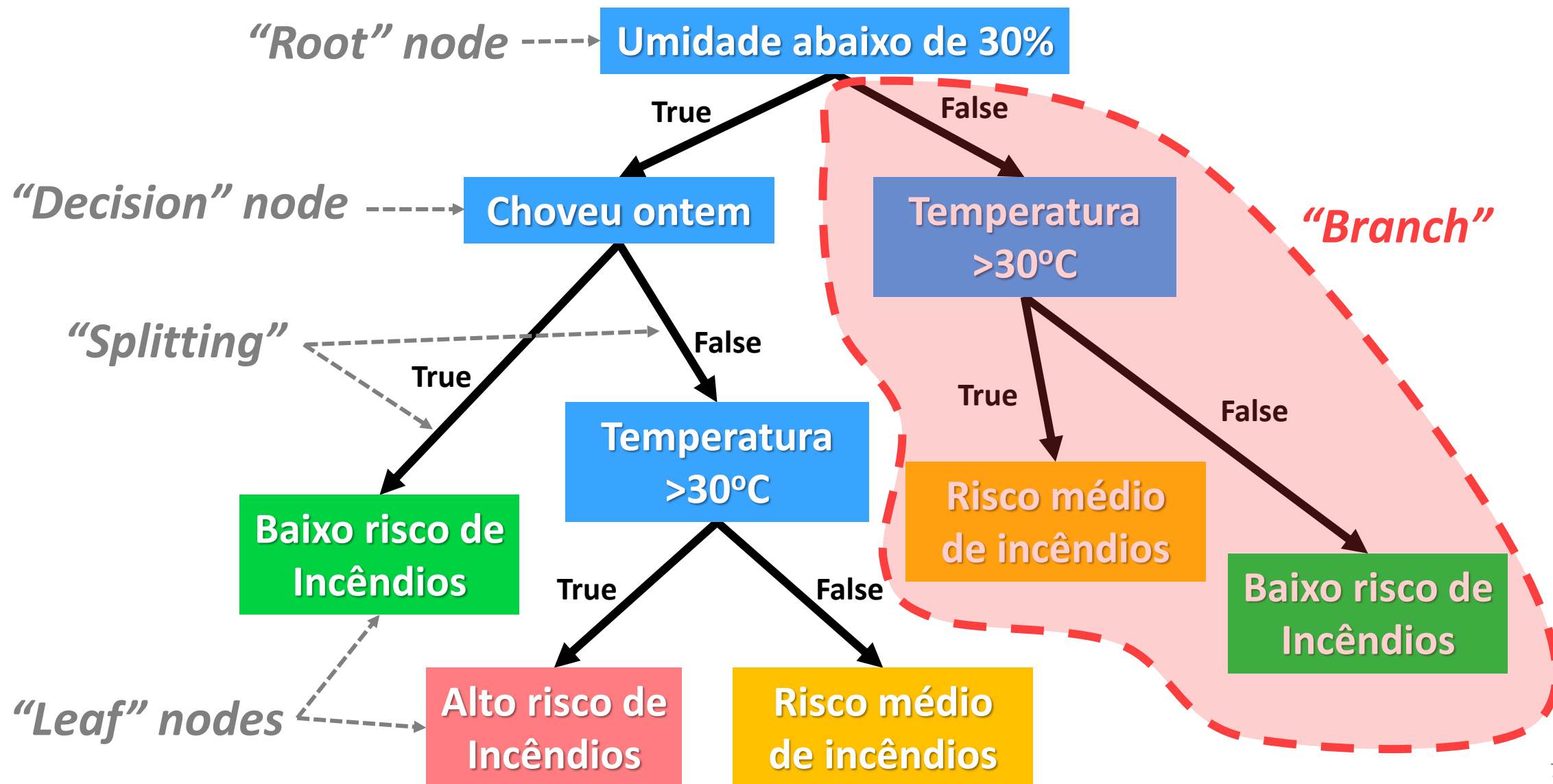
É uma árvore de regressão!



Esta árvore classifica dados em categorias...

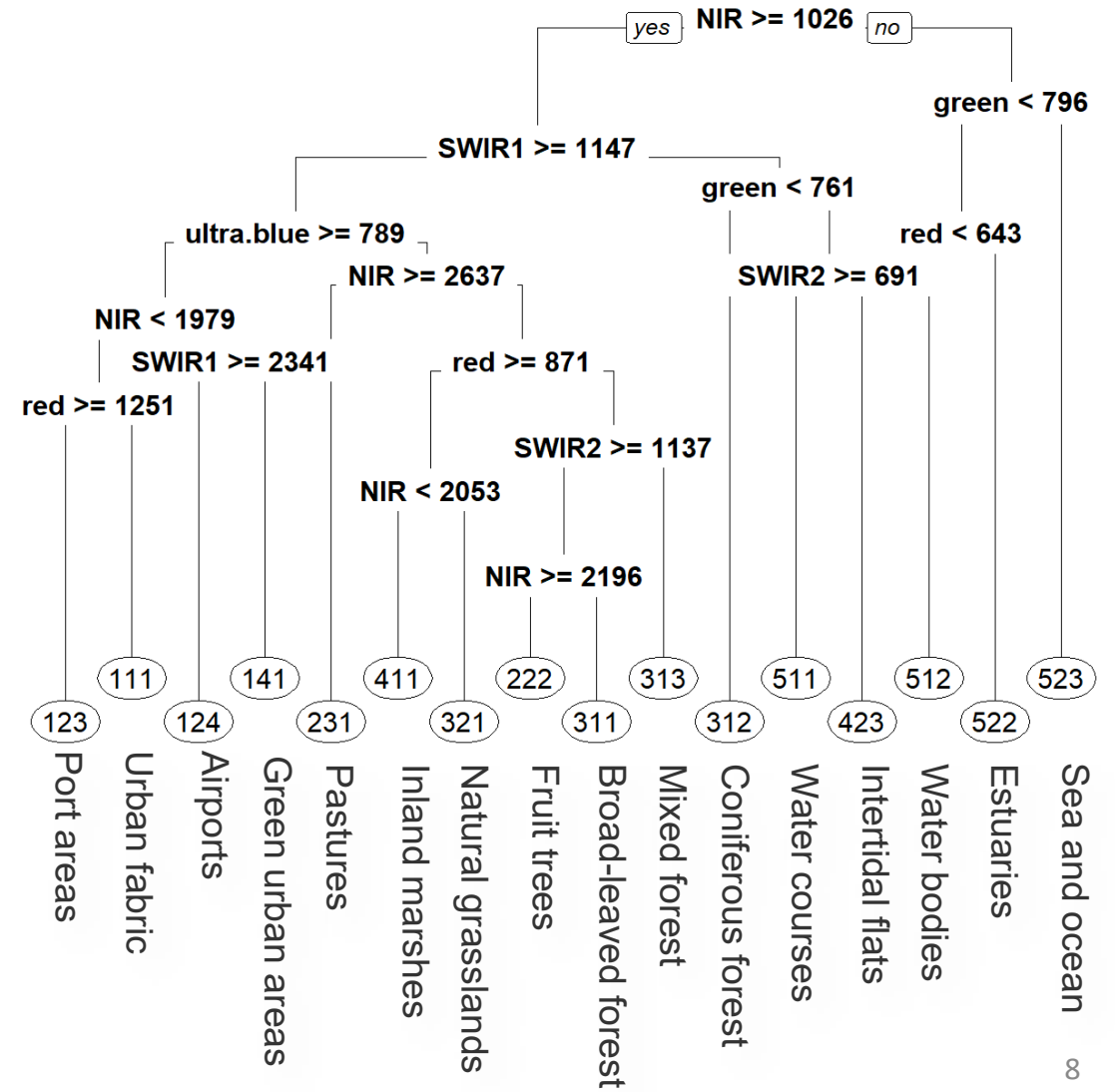
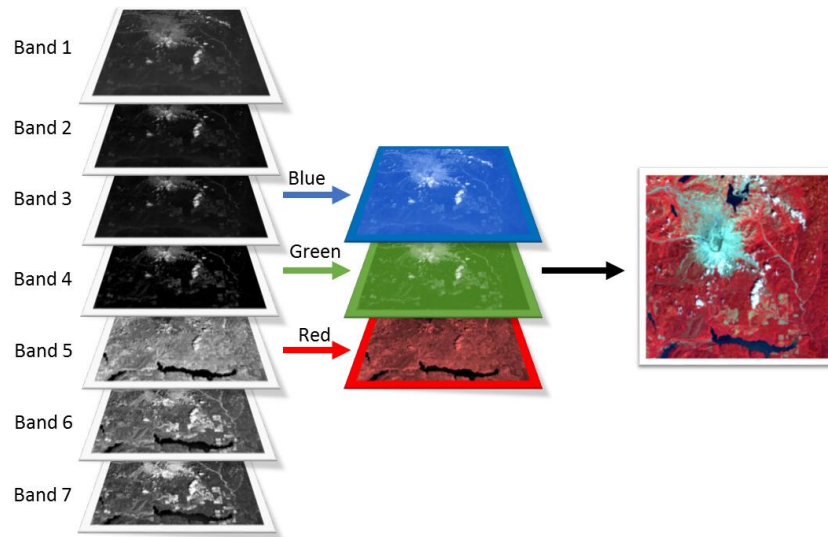
É uma Árvore de Classificação!

Árvores de decisão



Árvores de decisão

Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) Launched February 11, 2013	Bands
	Band 1 - Coastal aerosol
	Band 2 - Blue
	Band 3 - Green
	Band 4 - Red
	Band 5 - Near Infrared (NIR)
	Band 6 - SWIR 1
	Band 7 - SWIR 2
	Band 8 - Panchromatic
	Band 9 - Cirrus
	Band 10 - Thermal Infrared (TIRS) 1
Band 11 - Thermal Infrared (TIRS) 2	



Árvores de decisão

As árvores de decisão podem ser bastante complexas...

... mas eles são simples de interpretar



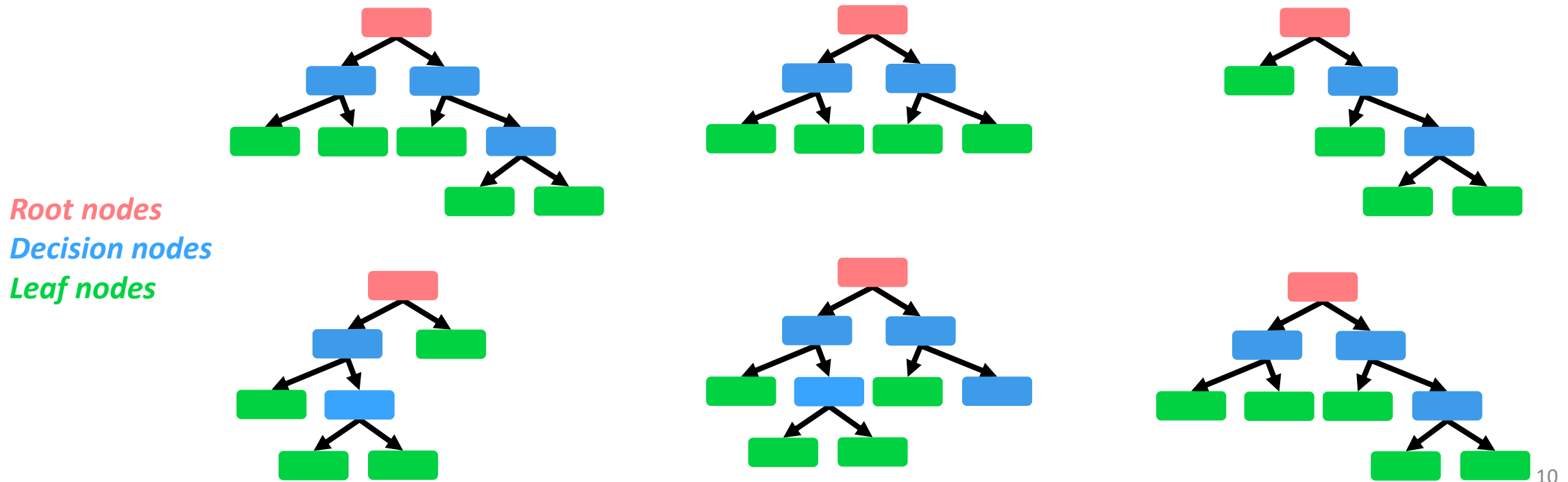
*“As árvores têm um aspecto que as impede de ser a ferramenta ideal para o aprendizado estatístico: a imprecisão” —The Elements of **Statistical Learning** (a.k.a. a “Bíblia” do Machine Learning)*

Elas funcionam bem com os dados usados para criá-las, mas não são muito flexíveis para lidar com novos dados

De uma árvore de decisão a uma floresta aleatória...

A vantagem das Florestas Aleatórias (Random Forests):

Simplicidade das árvores de decisão + flexibilidade = Maior precisão



Histórico...



© 2001 Kluwer Academic Publishers. Manufactured in The Netherlands.

Machine Learning, 45, 5–32, 2001

Random Forests

LEO BREIMAN

Statistics Department, University of California, Berkeley, CA 94720

Editor: Robert E. Schapire

Abstract. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Y. Freund & R. Schapire, *Machine Learning: Proceedings of the Thirteenth International conference*, * * *, 148–156), but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

Keywords: classification, regression, ensemble



Leo Breiman 1928-2005

[FOLLOW](#)

Professor of Statistics, [UC Berkeley](#).
Verified email at [stat.berkeley.edu](#) - [Homepage](#)

[Data Analysis](#) [Statistics](#) [Machine Learning](#)

[ARTICLES](#) [CITED BY](#)

TITLE	CITED BY	YEAR
Random forests	149207	2001
L Breiman Machine learning 45 (1), 5-32		

Aplicações de modelos de florestas aleatórias



Journal of Hazardous Materials
Volume 382, 15 January 2020, 120987



Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest

Open Access Technical Note

Application of the Random Forest Classifier to Map Irrigated Areas Using Google Earth Engine



Journal of Cleaner Production
Volume 260, 1 July 2020, 121082



Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability



Journal of Hydrology
Volume 589, October 2020, 125168



Research papers

Estimating annual runoff in response to forest change: A statistical method based on random forest



International Journal of Applied Earth Observation and Geoinformation
Volume 103, 1 December 2021, 102475



Identifying core driving factors of urban land use change from global land cover products and POI data using the random forest method



Journal of Cleaner Production
Volume 250, 20 March 2020, 119468



Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm



RESEARCH ARTICLE | Open Access | CC BY

MetaComNet: A random forest-based framework for making spatial predictions of plant–pollinator interactions



Ecological Modelling
Volume 419, 1 March 2020, 108947



Estimating dynamics of central hardwood forests using random forests



Applied Energy
Volume 305, 1 January 2022, 117916



A data-driven method of traffic emissions mapping with land use random forest models



Environmental Research
Volume 202, November 2021, 111660



Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Random Forests for Global and Regional Crop Yield Predictions

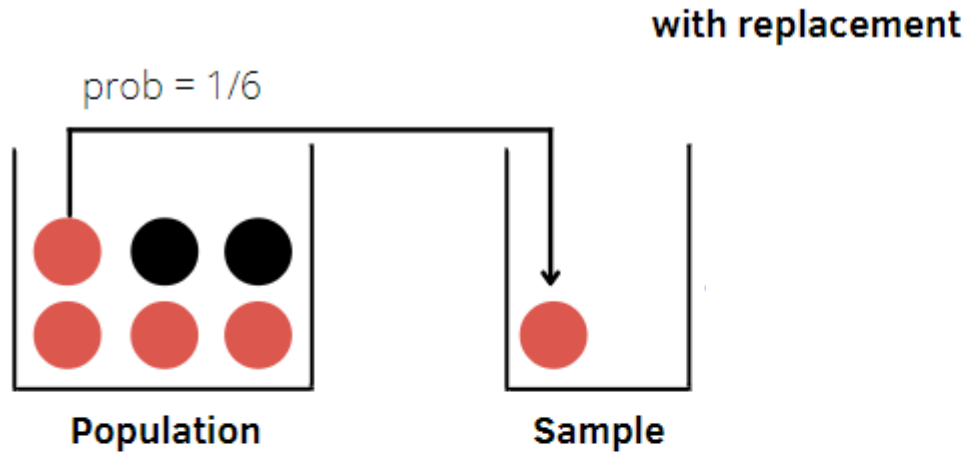
Agenda

- ❑ *Criando uma árvore de decisão com bootstrapping*
- ❑ *Usando um modelo de floresta aleatória*
- ❑ *Construindo um modelo de Floresta Aleatória*
- ❑ *Florestas aleatórias de regressão*

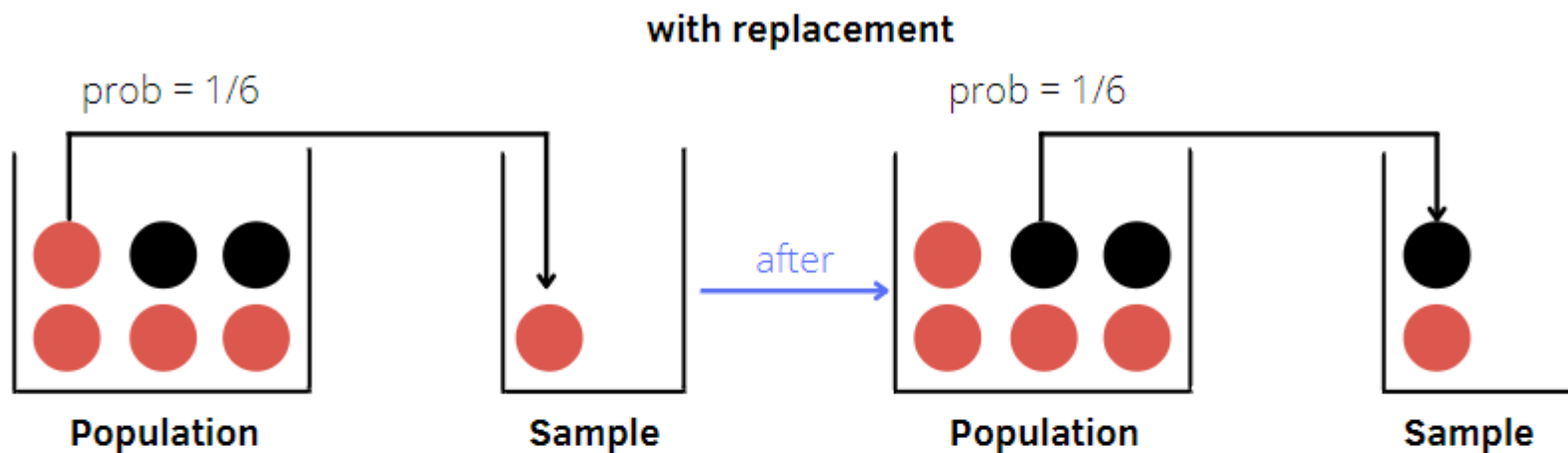
Agenda

- ❑ ***Criando uma árvore de decisão com bootstrapping***
- ❑ *Usando um modelo de floresta aleatória*
- ❑ *Construindo um modelo de Floresta Aleatória*
- ❑ *Florestas aleatórias de regressão*

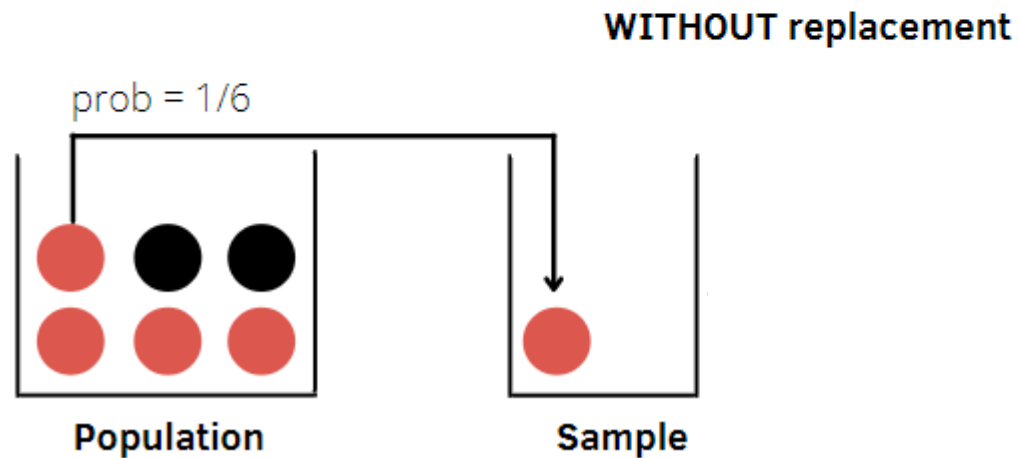
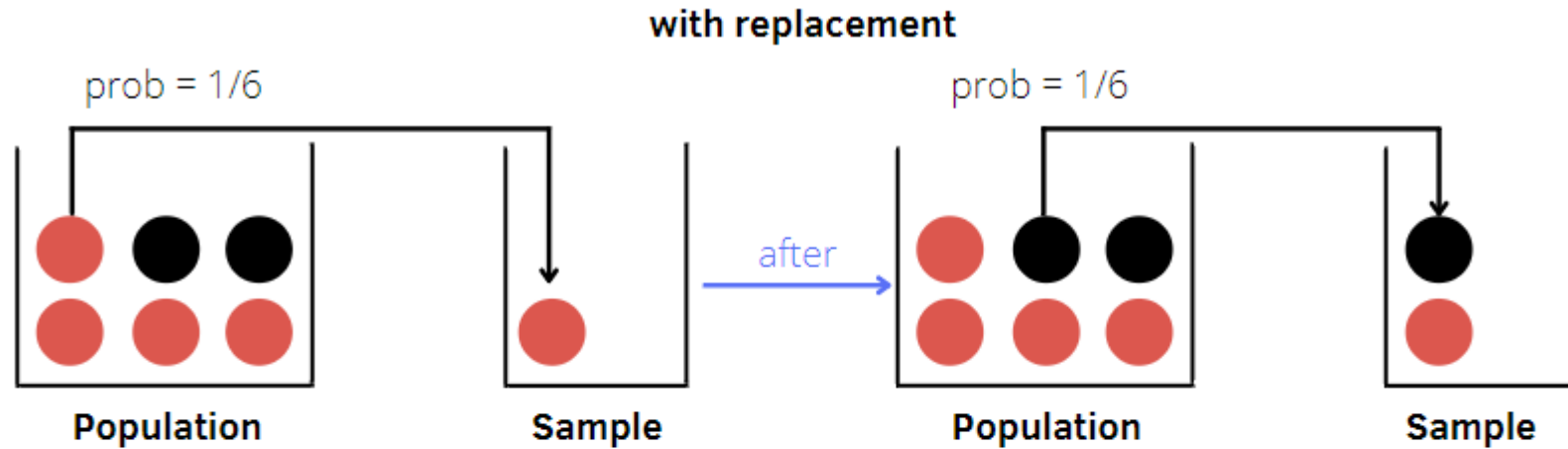
Amostragem: com & sem substituição



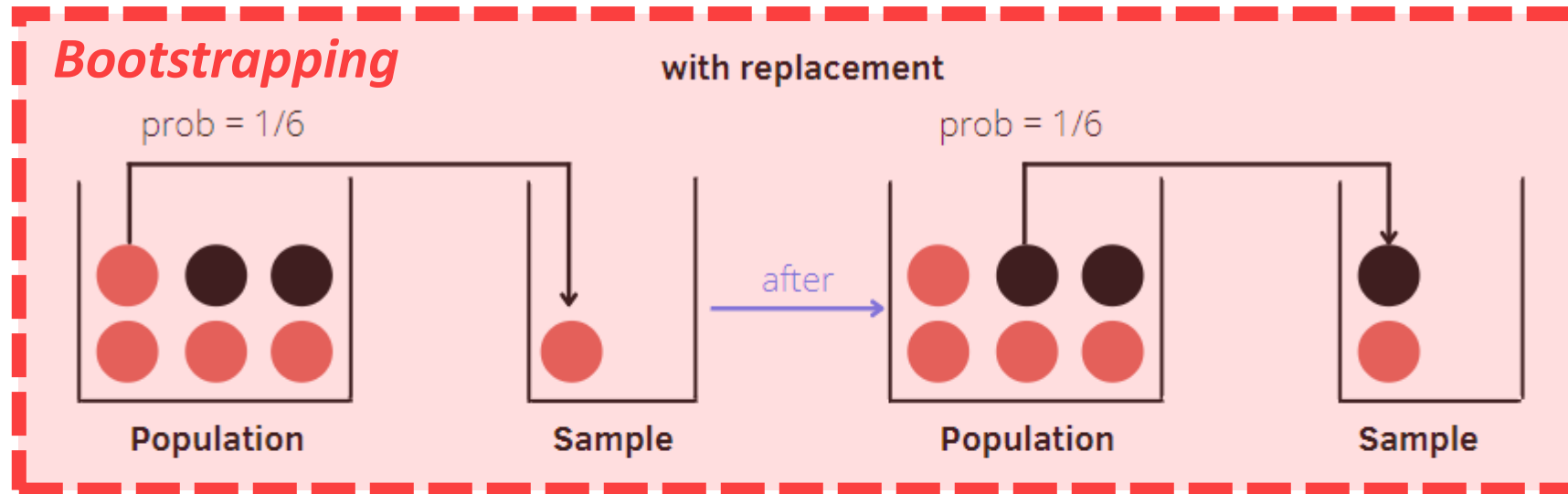
Amostragem: com & sem substituição



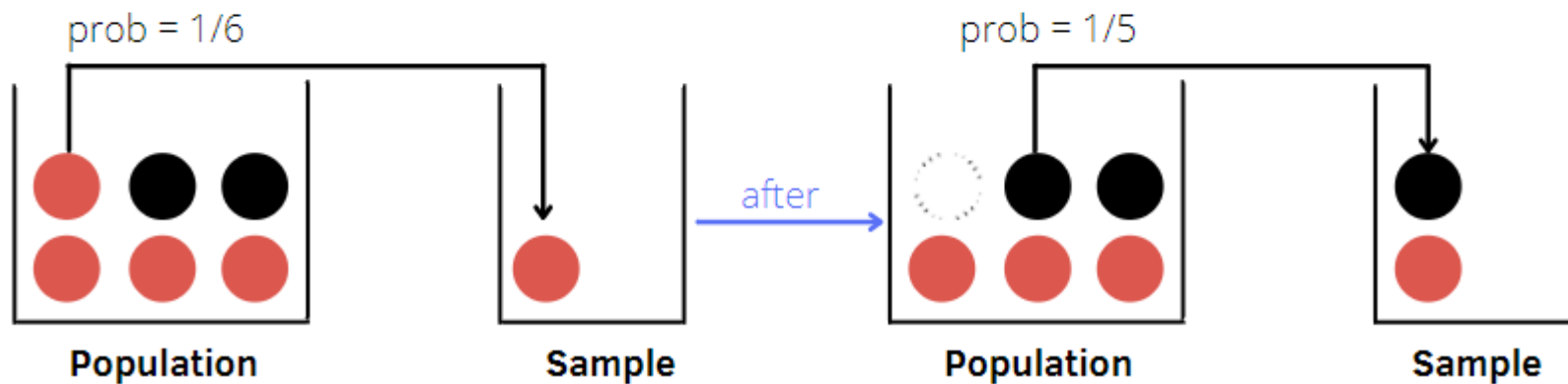
Amostragem: com & sem substituição



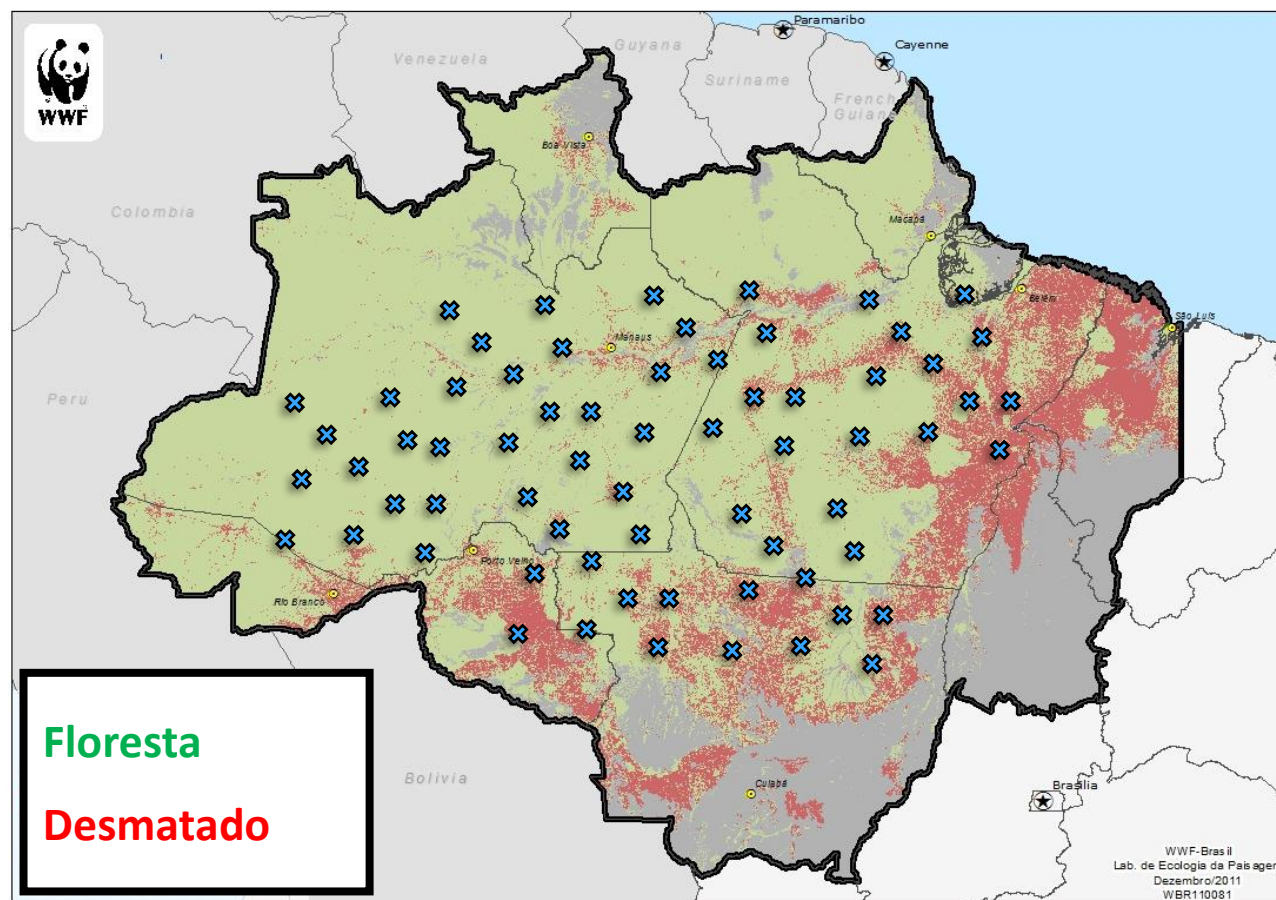
Amostragem: com & sem substituição



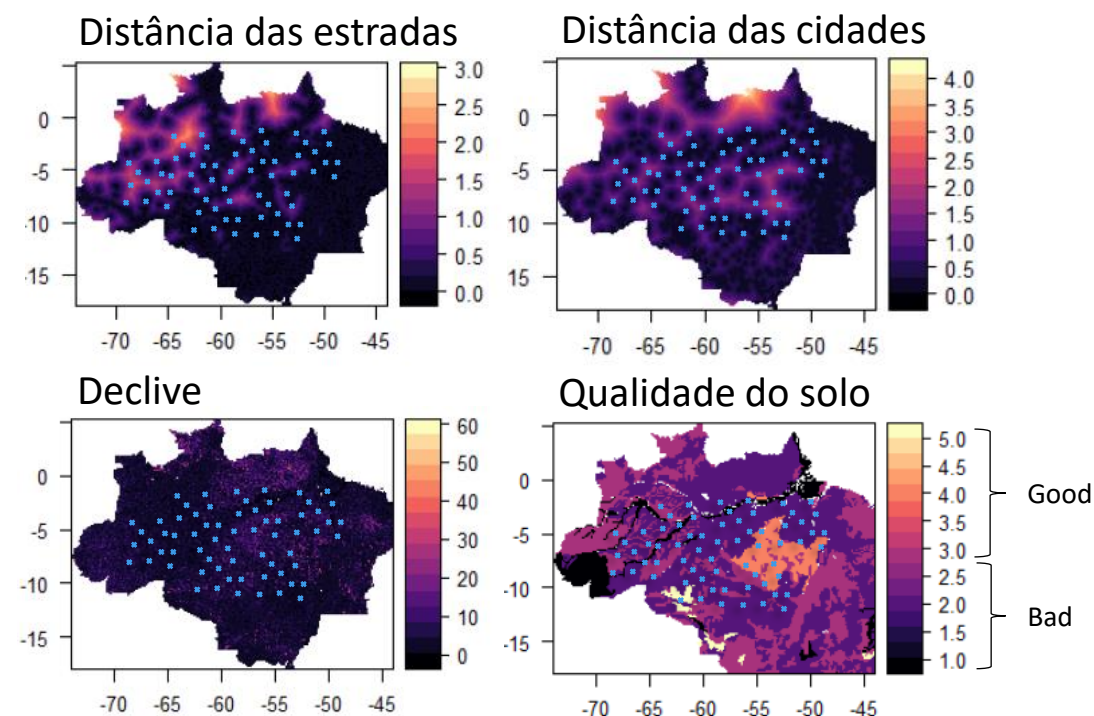
WITHOUT replacement



Exemplo: Previsão de desmatamento



Amostra 50 pontos: Desmatado (sim/não)



Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
-----------	------------	------------	-------	------------	-------------

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...					
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...					
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...					
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No
2	25.6	14.1	0.5	Bad	No

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...					
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...					
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...					
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
18	24.6	13.1	5.6	Bad	Yes

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
18	24.6	13.1	5.6	Bad	Yes
1	21.4	3.2	12	Good	Yes

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
18	24.6	13.1	5.6	Bad	Yes
1	21.4	3.2	12	Good	Yes
...	...	...	...	...	...

Bootstrapping

Dados originais (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
1	21.4	3.2	12	Good	Yes
2	25.6	14.1	0.5	Bad	No
3	18.9	3.5	2.3	Good	No
4	7.8	5.6	11	Good	Yes
5	6.9	2.9	8	Bad	No
6	7.2	4.5	7	Good	Yes
...	...	...	...	...	...
18	24.6	13.1	5.6	Bad	Yes
19	26.2	6.6	8	Bad	No
20	18.9	4.5	12.1	Good	Yes
21	6.8	2.3	13	Bad	No
22	18	4.7	9.8	Bad	No
...	...	...	...	...	...
33	48.1	13	3	Bad	No
34	23	5.9	3.5	Good	Yes
36	44.7	22.3	2.1	Bad	No
...	...	...	...	...	...
49	5.8	2.3	0.8	Bad	No
50	9.8	1.1	0.4	Good	Yes

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
18	24.6	13.1	5.6	Bad	Yes
1	21.4	3.2	12	Good	Yes
...	...	...	...	...	...
15	1.4	3.8	2	Bad	Yes
39	41.4	31.2	12	Bad	No
3	18.9	3.5	2.3	Good	No
33	48.1	13	3	Bad	No

Bootstrapping

Agora, criaremos uma árvore de decisão com base no conjunto de ***dados bootstrapped***...

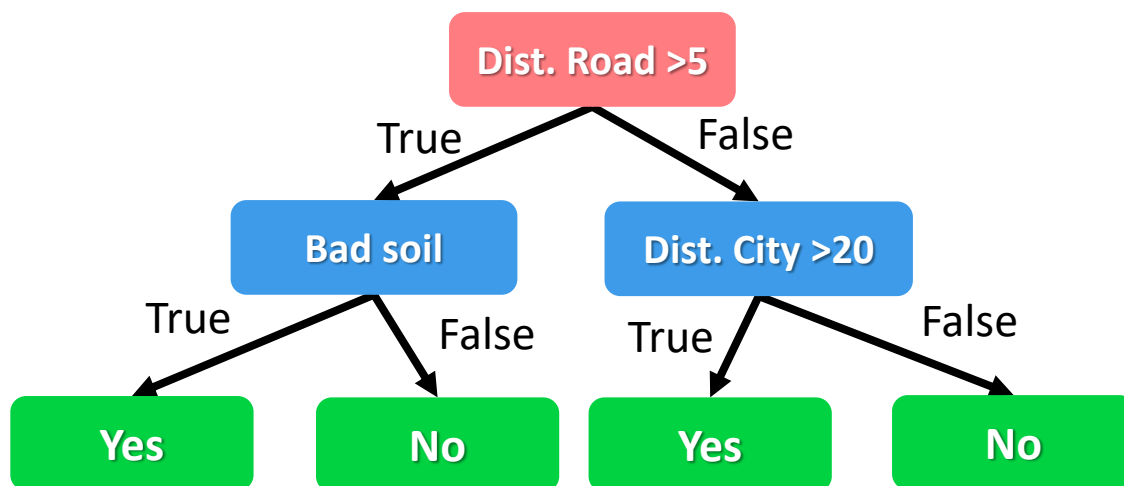
Normalmente, apenas algumas variáveis são (aleatoriamente) consideradas na divisão de cada ***decision node***

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
33	48.1	13	3	Bad	No
18	24.6	13.1	5.6	Bad	Yes
2	25.6	14.1	0.5	Bad	No
6	7.2	4.5	7	Good	Yes
19	26.2	6.6	8	Bad	No
36	44.7	22.3	2.1	Bad	No
33	48.1	13	3	Bad	No
22	18	4.7	9.8	Bad	No
50	9.8	1.1	0.4	Good	Yes
36	44.7	22.3	2.1	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
18	24.6	13.1	5.6	Bad	Yes
1	21.4	3.2	12	Good	Yes
...	...	...	...	...	...
15	1.4	3.8	2	Bad	Yes
39	41.4	31.2	12	Bad	No
3	18.9	3.5	2.3	Good	No
33	48.1	13	3	Bad	No

Do bootstrapping à Árvores de Decisão

*E repetimos esse processo, criando a cada vez um conjunto de **dados bootstrapped** diferente e uma nova árvore de decisão ...*



Podemos fazer isso centenas de vezes! Essa variedade de árvores é o que torna Florestas Aleatórias mais eficazes do que as árvores de decisão individuais!

Dados bootstrapped (n = 50)

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
20	18.9	4.5	12.1	Good	Yes
3	18.9	3.5	2.3	Good	No
15	1.4	3.8	2	Bad	Yes
21	6.8	2.3	13	Bad	No
50	9.8	1.1	0.4	Good	Yes
1	21.4	3.2	12	Good	Yes
34	23	5.9	3.5	Good	Yes
22	18	4.7	9.8	Bad	No
49	5.8	2.3	0.8	Bad	No
21	6.8	2.3	13	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
7	2.6	17.1	1.6	Good	Yes
50	9.8	1.1	0.4	Good	Yes
...	...	...	...	...	...
34	23	5.9	3.5	Good	Yes
21	6.8	2.3	13	Bad	No
46	17.2	24.5	17	Good	No
4	7.8	5.6	11	Good	Yes

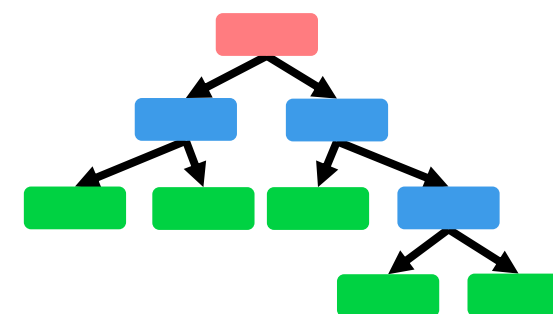
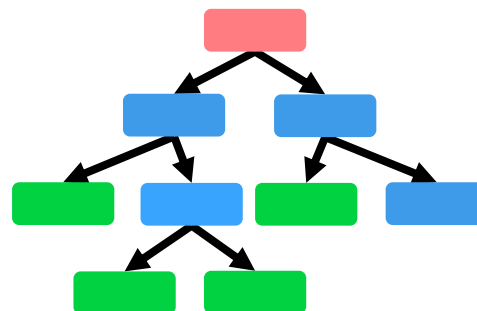
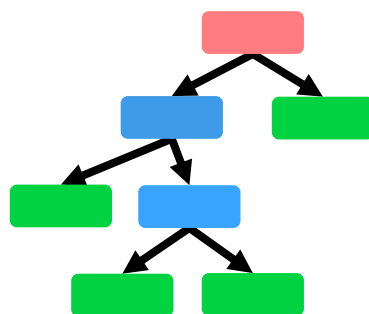
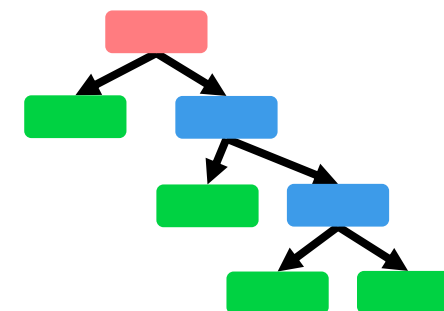
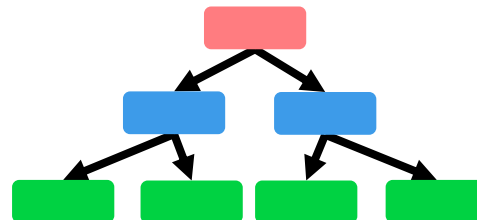
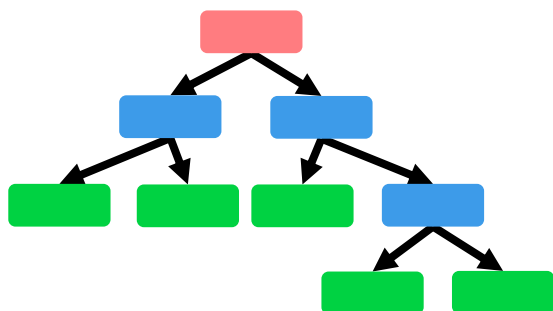
Agenda

- ☒ *Criando uma árvore de decisão com bootstrapping*
- ☐ ***Usando um modelo de floresta aleatória***
- ☐ *Construindo um modelo de Floresta Aleatória*
- ☐ *Florestas aleatórias de regressão*

Como usamos uma Floresta Aleatória?

Vamos supor que nossa Floresta Aleatória utilizada para prever “desmatamento” tenha apenas 6 árvores...

Root nodes
Decision nodes
Leaf nodes



Como usamos uma Floresta Aleatória?

Primeiro, olhamos para um novo local/pixel...

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
"New"	8.9	14	22.1	Good	???

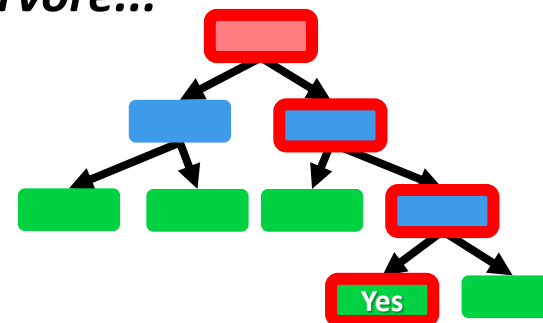
Estes nós sabemos ...

Is this location likely to
undergo forest loss?

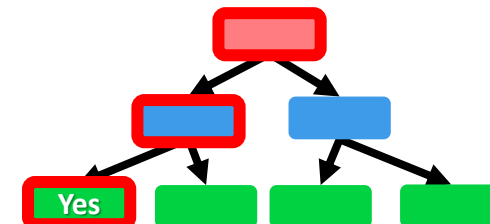
Yes = 5

No = 1

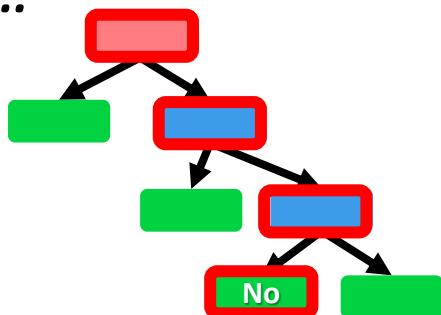
Passamos os dados desse novo pixel pela primeira árvore...



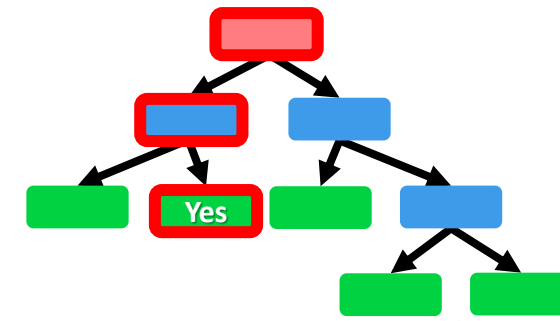
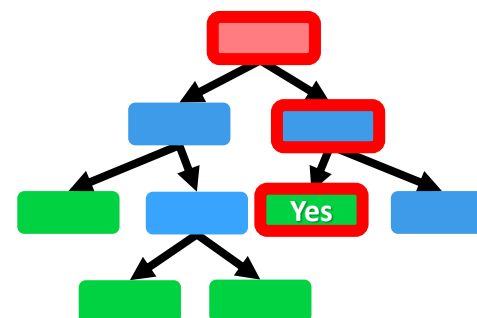
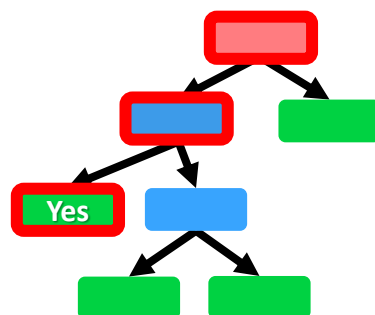
... e através da segunda árvore...



... e através da terceira árvore...



... e assim por diante...



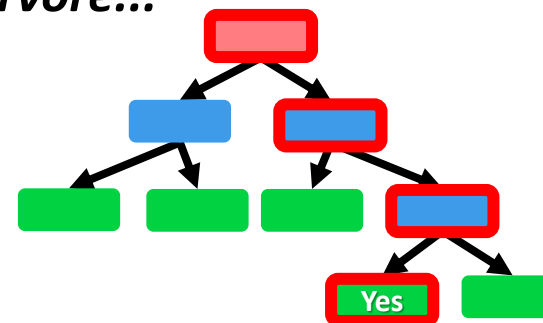
Como usamos uma Floresta Aleatória?

Fazer o **bootstrapping** e **agregar** os resultados para fazer uma previsão é chamado de “**bagging**”

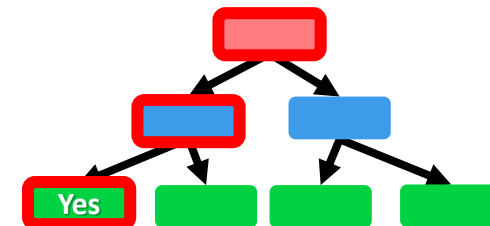
... ajuda a evitar o overfitting

Is this location likely to undergo forest loss?
Yes = 5
No = 1

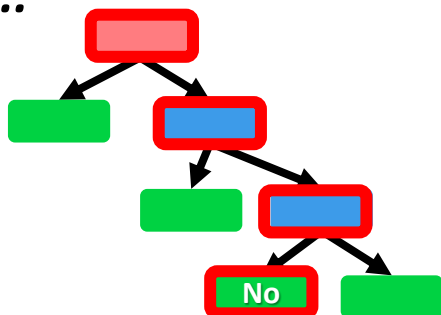
Passamos os dados desse novo pixel pela primeira árvore...



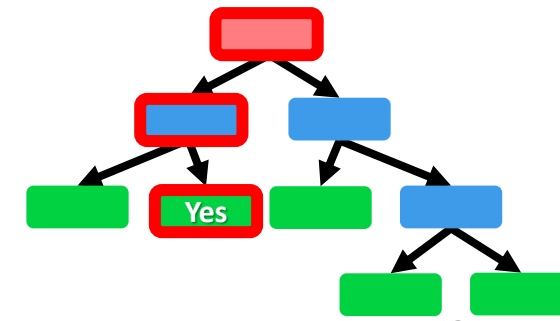
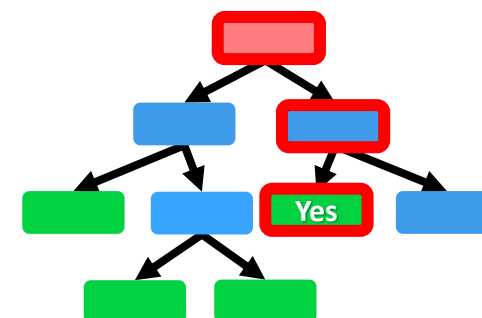
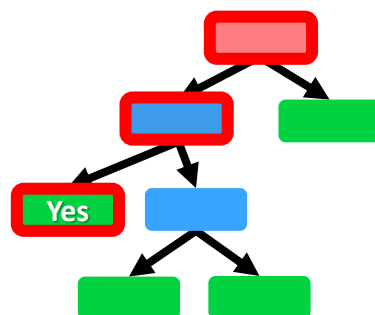
... e através da segunda árvore...



... e através da terceira árvore...



... e assim por diante...



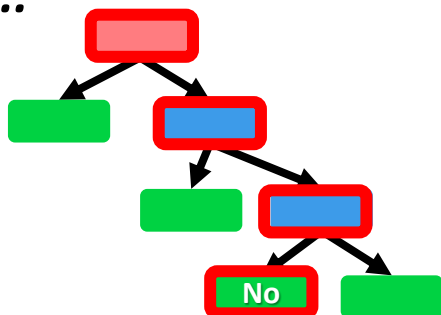
Como usamos uma Floresta Aleatória?

E então temos uma previsão...

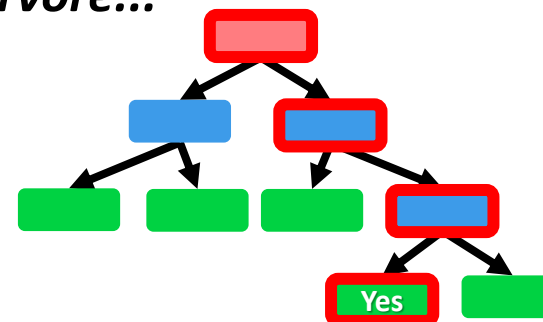
Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
"New"	8.9	14	22.1	Good	Yes (5/6)

Is this location likely to
undergo forest loss?
Yes = 5
No = 1

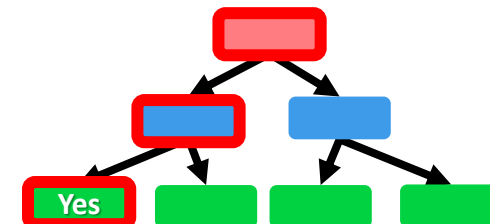
*... e através da terceira
árvore...*



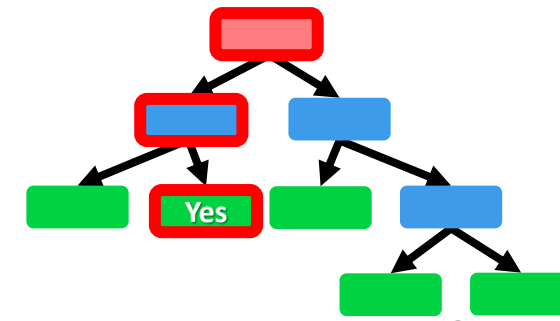
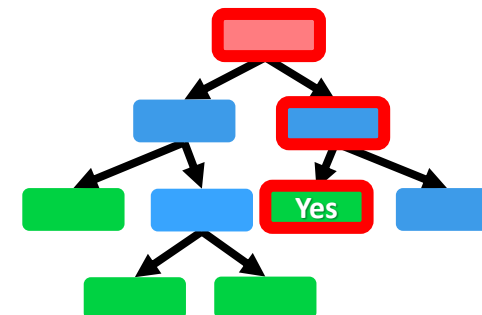
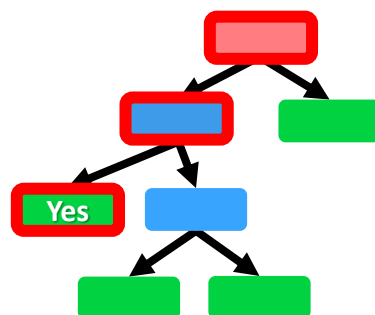
*Passamos os dados desse
novo pixel pela primeira
árvore...*



*... e através da segunda
árvore...*



... e assim por diante...

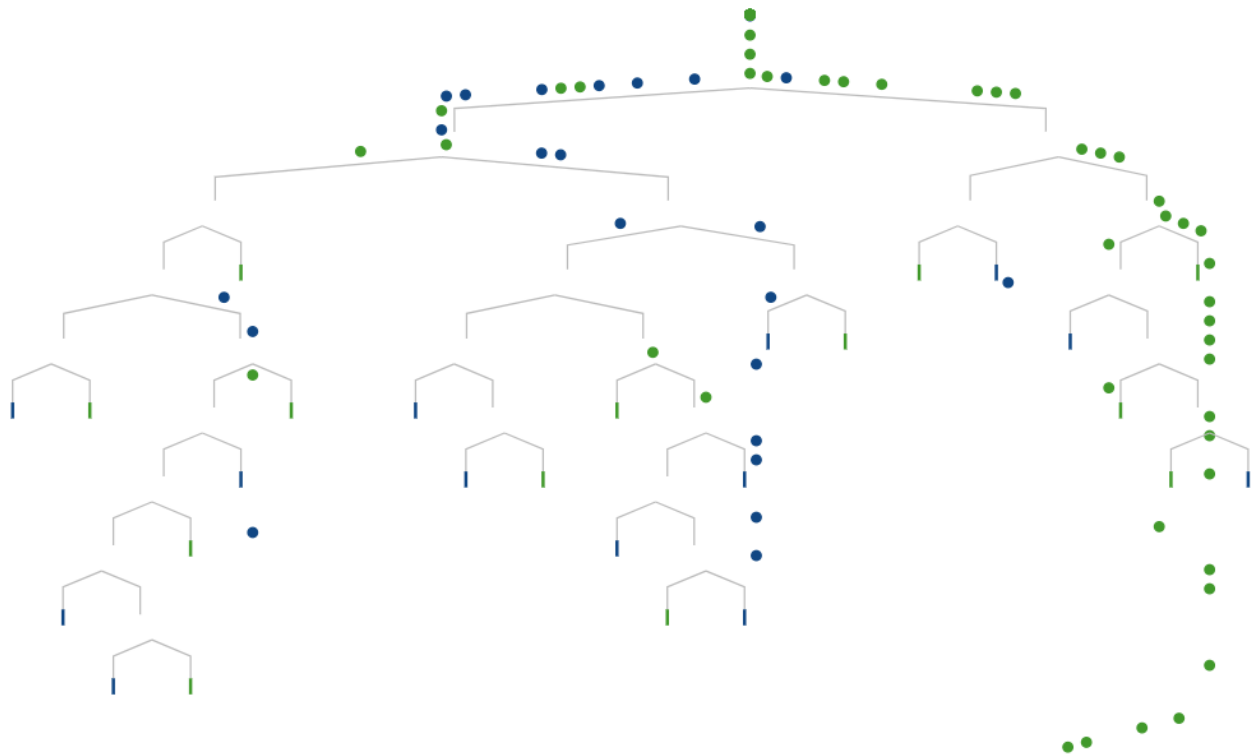


Exemplo...

<http://www.r2d3.us/uma-introducao-visual-ao-aprendizado-de-maquina-1/>

Fazendo previsões

O modelo de árvore de decisão recém treinado determina se um imóvel está localizado em São Francisco ou em Nova Iorque, executando cada ponto de dados por meio de ramificações.



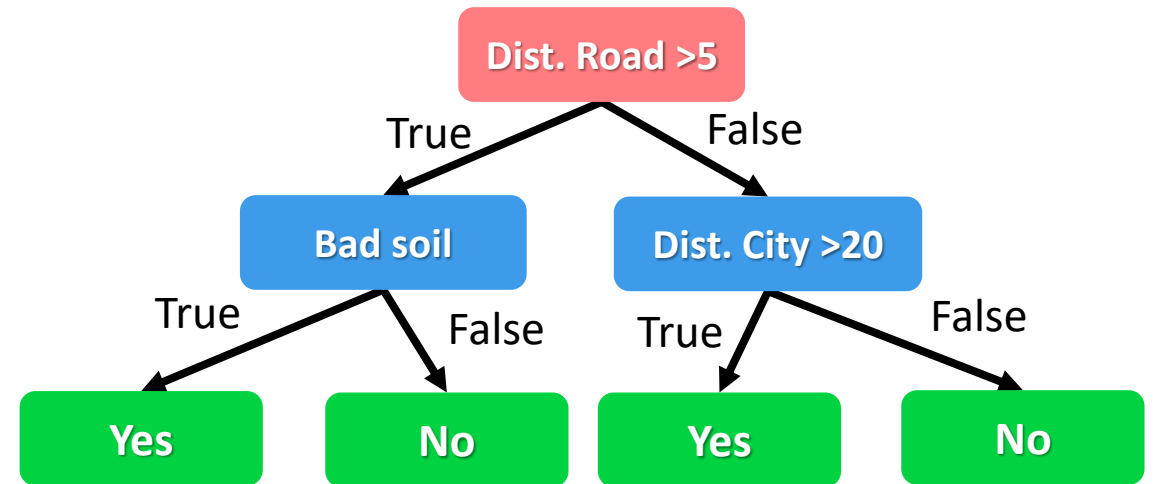
Agenda

- ☒ *Criando uma árvore de decisão com bootstrapping*
- ☒ *Usando um modelo de floresta aleatória*
- ☐ ***Construindo um modelo de Floresta Aleatória***
- ☐ *Florestas aleatórias de regressão*

Como as Árvores de Decisão são construídas?

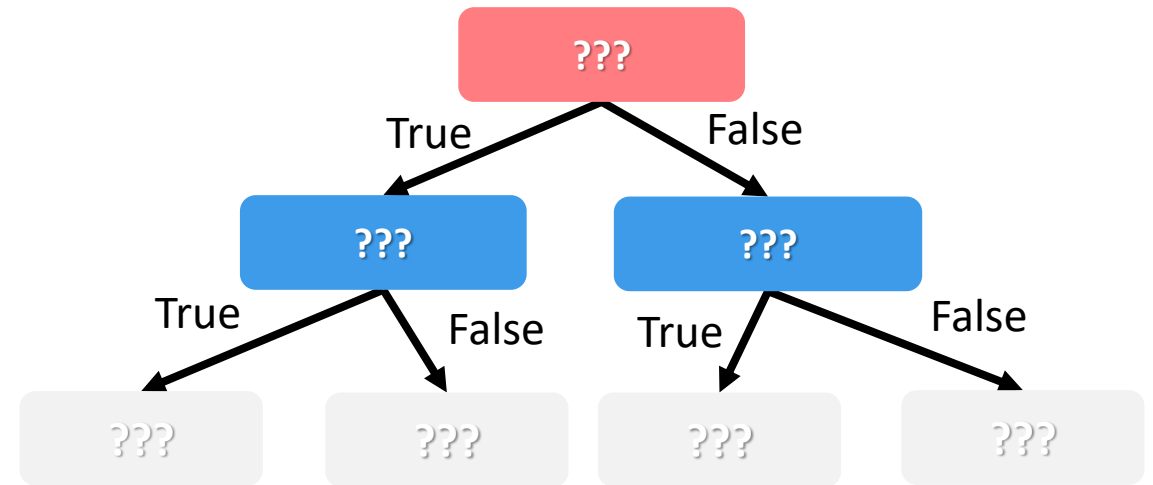
Dados bootstrapped

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
20	18.9	4.5	12.1	Good	Yes
3	18.9	3.5	2.3	Good	No
15	1.4	3.8	2	Bad	Yes
21	6.8	2.3	13	Bad	No
50	9.8	1.1	0.4	Good	Yes
1	21.4	3.2	12	Good	Yes
34	23	5.9	3.5	Good	Yes
22	18	4.7	9.8	Bad	No
49	5.8	2.3	0.8	Bad	No
21	6.8	2.3	13	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
7	2.6	17.1	1.6	Good	Yes
50	9.8	1.1	0.4	Good	Yes
...	...	...	...	...	...
34	23	5.9	3.5	Good	Yes
21	6.8	2.3	13	Bad	No
46	17.2	24.5	17	Good	No
4	7.8	5.6	11	Good	Yes



Como as Árvores de Decisão são construídas?

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
20	18.9	4.5	12.1	Good	Yes
3	18.9	3.5	2.3	Good	No
15	1.4	3.8	2	Bad	Yes
21	6.8	2.3	13	Bad	No
50	9.8	1.1	0.4	Good	Yes
1	21.4	3.2	12	Good	Yes
34	23	5.9	3.5	Good	Yes
22	18	4.7	9.8	Bad	No
49	5.8	2.3	0.8	Bad	No
21	6.8	2.3	13	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
7	2.6	17.1	1.6	Good	Yes
50	9.8	1.1	0.4	Good	Yes
...	...	...	...	...	...
34	23	5.9	3.5	Good	Yes
21	6.8	2.3	13	Bad	No
46	17.2	24.5	17	Good	No
4	7.8	5.6	11	Good	Yes



Como as Árvores de Decisão são construídas?

Observe que neste exemplo estamos usando todas as variáveis independentes, mas nem sempre é esse o caso...

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
20	18.9	4.5	12.1	Good	Yes
3	18.9	3.5	2.3	Good	No
15	1.4	3.8	2	Bad	Yes
21	6.8	2.3	13	Bad	No
50	9.8	1.1	0.4	Good	Yes
1	21.4	3.2	12	Good	Yes
34	23	5.9	3.5	Good	Yes
22	18	4.7	9.8	Bad	No
49	5.8	2.3	0.8	Bad	No
21	6.8	2.3	13	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
7	2.6	17.1	1.6	Good	Yes
50	9.8	1.1	0.4	Good	Yes
...	...	...	...	...	...
34	23	5.9	3.5	Good	Yes
21	6.8	2.3	13	Bad	No
46	17.2	24.5	17	Good	No
4	7.8	5.6	11	Good	Yes

Primeiro, o algoritmo encontra o limite ideal para cada variável contínua (Dist. City, Dist. Road e Slope) que melhor explica Forest loss (=melhor separa 'yes' dos 'no')

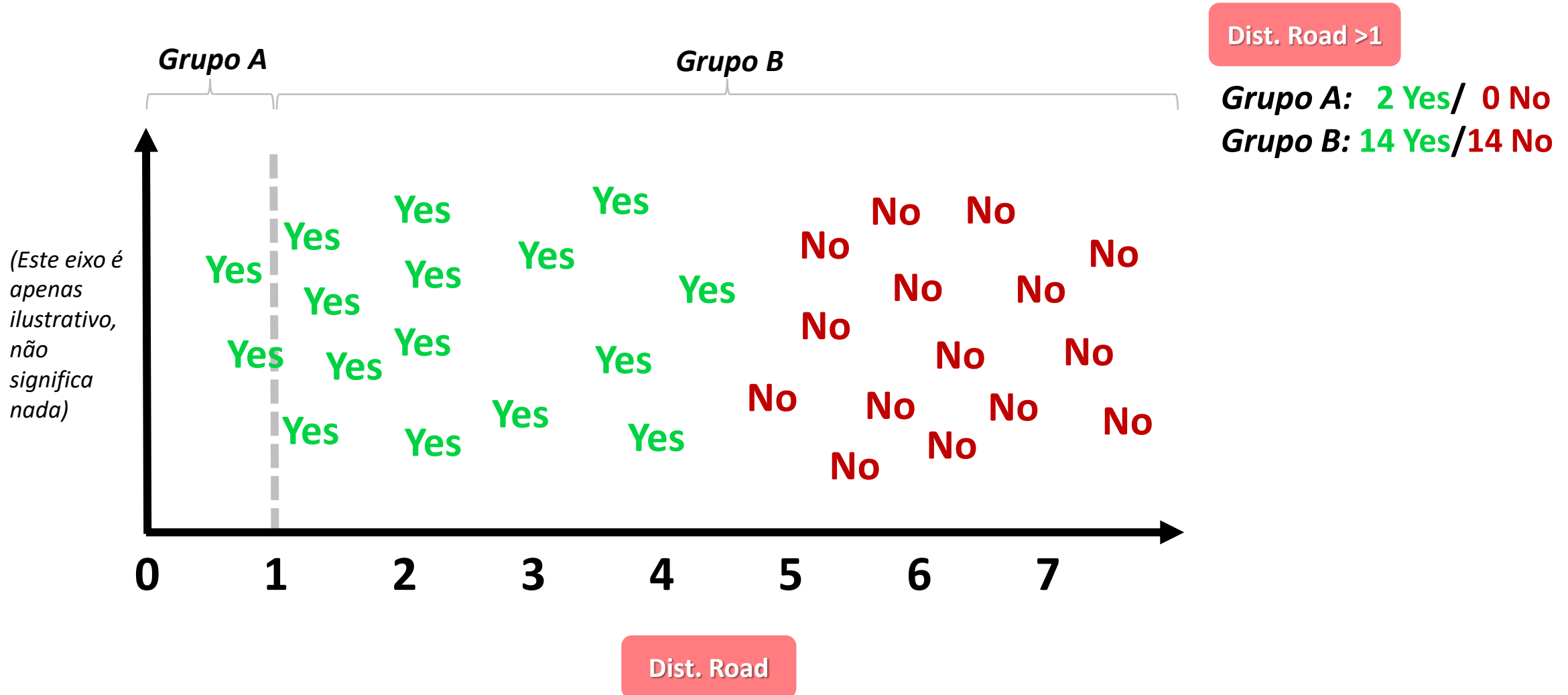
Por exemplo:

Dist. Road >5

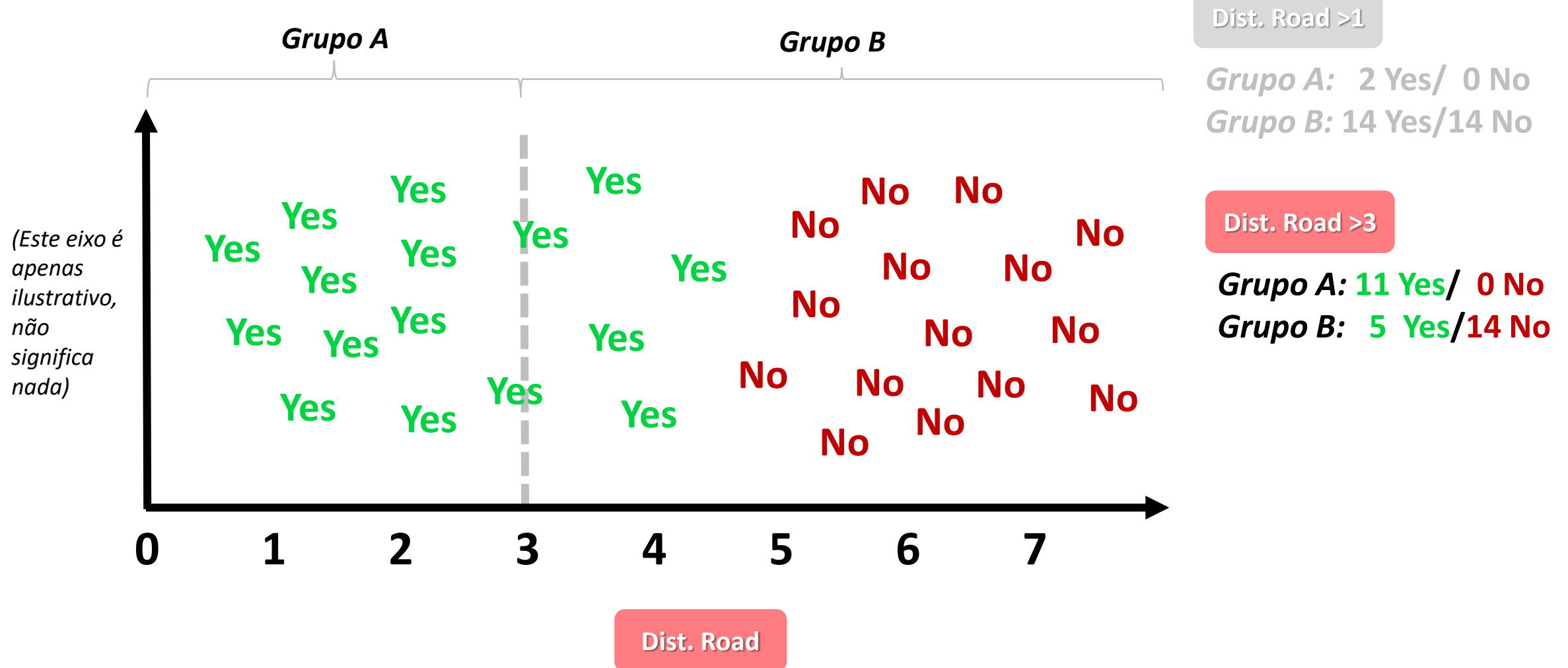
Dist. City <9

Slope <15

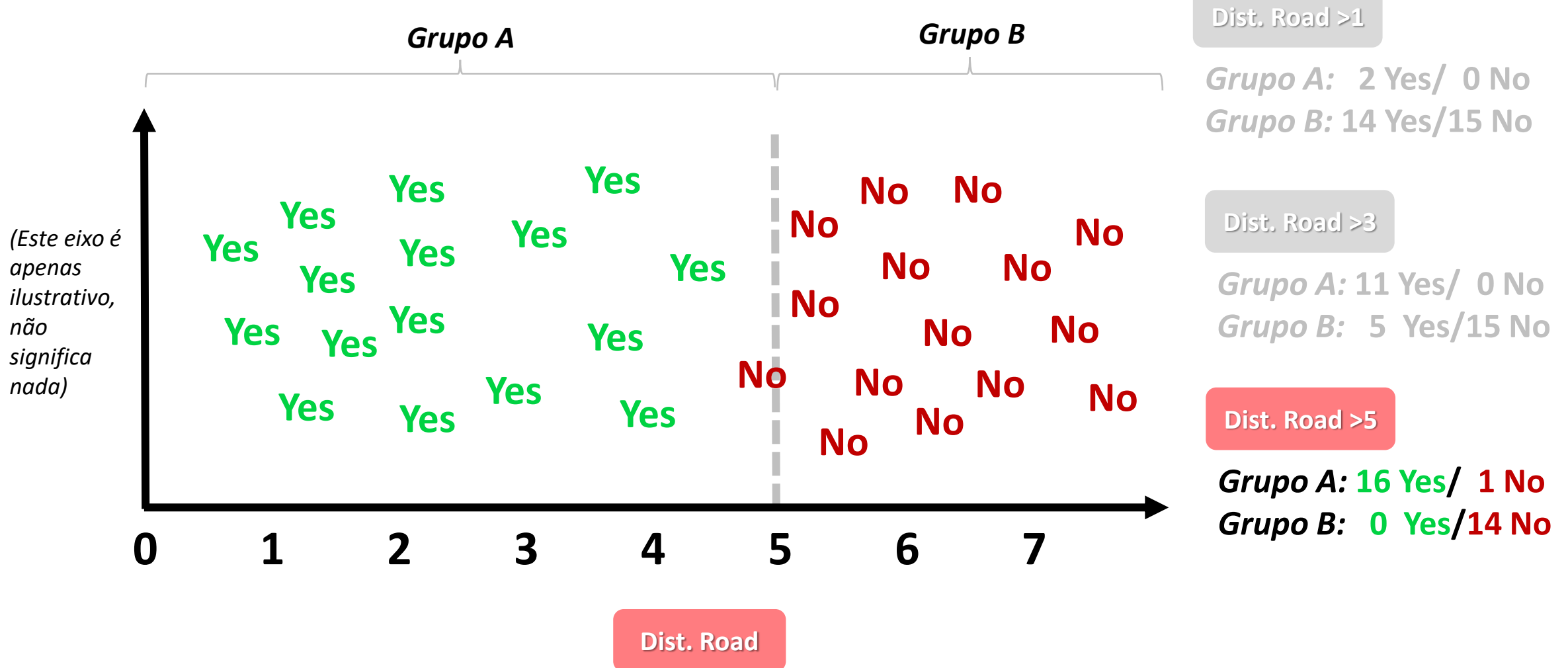
Como as Árvores de Decisão são construídas?



How are the Decision Trees constructed?



How are the Decision Trees constructed?



Como as Árvores de Decisão são construídas?

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
20	18.9	4.5	12.1	Good	Yes
3	18.9	3.5	2.3	Good	No
15	1.4	3.8	2	Bad	Yes
21	6.8	2.3	13	Bad	No
50	9.8	1.1	0.4	Good	Yes
1	21.4	3.2	12	Good	Yes
34	23	5.9	3.5	Good	Yes
22	18	4.7	9.8	Bad	No
49	5.8	2.3	0.8	Bad	No
21	6.8	2.3	13	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
7	2.6	17.1	1.6	Good	Yes
50	9.8	1.1	0.4	Good	Yes
...	...	...	...	...	...
34	23	5.9	3.5	Good	Yes
21	6.8	2.3	13	Bad	No
46	17.2	24.5	17	Good	No
4	7.8	5.6	11	Good	Yes

Primeiro, o algoritmo encontra o limite ideal para cada variável contínua (Dist. City, Dist. Road e Slope) que melhor explica Forest loss (=melhor separa 'yes' dos 'no')

Por exemplo:

Dist. Road >5

Dist. City <9

Slope <15

Em seguida, o algoritmo identifica quais das variáveis contínuas (com seus limites ótimos), incluindo as variáveis categóricas (Soil qual.) melhor explica Forest loss

Como as Árvores de Decisão são construídas?

Vamos supor que esses sejam os limites ideais para separar os dados entre grupos de **Yes** e **No** usando essas **variáveis**:

Dist. Road >1

Grupo A: 2 Yes/ 0 No
Grupo B: 14 Yes/15 No

Dist. Road >3

Grupo A: 11 Yes/ 0 No
Grupo B: 5 Yes/15 No

Good/bad soil

Grupo A: 10 Yes/ 7 No
Grupo B: 1 Yes/13 No

Slope <15

Grupo A: 11 Yes/ 6 No
Grupo B: 3 Yes/11 No

Dist. City <9

Grupo A: 14 Yes/ 3 No
Grupo B: 2 Yes/12 No

Dist. Road >5

Grupo A: 16 Yes/ 1 No
Grupo B: 0 Yes/14 No

Qual dessas variáveis fez o melhor trabalho dividindo os grupos **Yes** e **No**? Em outras palavras, qual delas mais se aproximou de um resultado com um grupo de **100% Yes/0% No** e outro de **0% Yes/100% No**?

Como as Árvores de Decisão são construídas?

Embora essa seja a ideia geral, há cálculos envolvidos nesta etapa!

Vamos supor que esses sejam os limites

ideais para separar os dados entre grupos

Esses cálculos estimam a "pureza" das divisões (ou seja, 100% Sim/0% Não e vice-versa): o objetivo é maximizar a pureza (ou minimizar a

impureza)

Good/bad soil

Slope <15

Dist. City <9

Dist. Road >5

Grupo A: 10 Yes/ 7 No

Grupo A: 11 Yes/ 6 No

Grupo A: 14 Yes/ 3 No

Grupo A: 16 Yes/ 1 No

Existem diferentes métodos/métricas para medir a "pureza" das divisões, como Impureza de Gini, Entropia e Information Gain

Qual dessas variáveis fez o melhor trabalho dividindo os grupos Yes e No? Em outras palavras, qual delas mais se aproximou de um resultado com um grupo de 100% Yes/0% No e outro de 0% Yes/100% No?

Como as Árvores de Decisão são construídas?

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
20	18.9	4.5	12.1	Good	Yes
3	18.9	3.5	2.3	Good	No
15	1.4	3.8	2	Bad	Yes
21	6.8	2.3	13	Bad	No
50	9.8	1.1	0.4	Good	Yes
1	21.4	3.2	12	Good	Yes
34	23	5.9	3.5	Good	Yes
22	18	4.7	9.8	Bad	No
49	5.8	2.3	0.8	Bad	No
21	6.8	2.3	13	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
7	2.6	17.1	1.6	Good	Yes
50	9.8	1.1	0.4	Good	Yes
...	...	...	...	...	...
34	23	5.9	3.5	Good	Yes
21	6.8	2.3	13	Bad	No
46	17.2	24.5	17	Good	No
4	7.8	5.6	11	Good	Yes

Primeiro, o algoritmo encontra o limite ideal para cada variável contínua (Dist. City, Dist. Road e Slope) que melhor explica Forest loss (=melhor separa 'yes' dos 'no')

Por exemplo:

Dist. Road >5

Dist. City <9

Slope <15

Em seguida, o algoritmo identifica quais das variáveis contínuas (com seus limites ótimos), incluindo as variáveis categóricas (Soil qual.) melhor explica Forest loss

Dist. Road >5

Dist. City <9

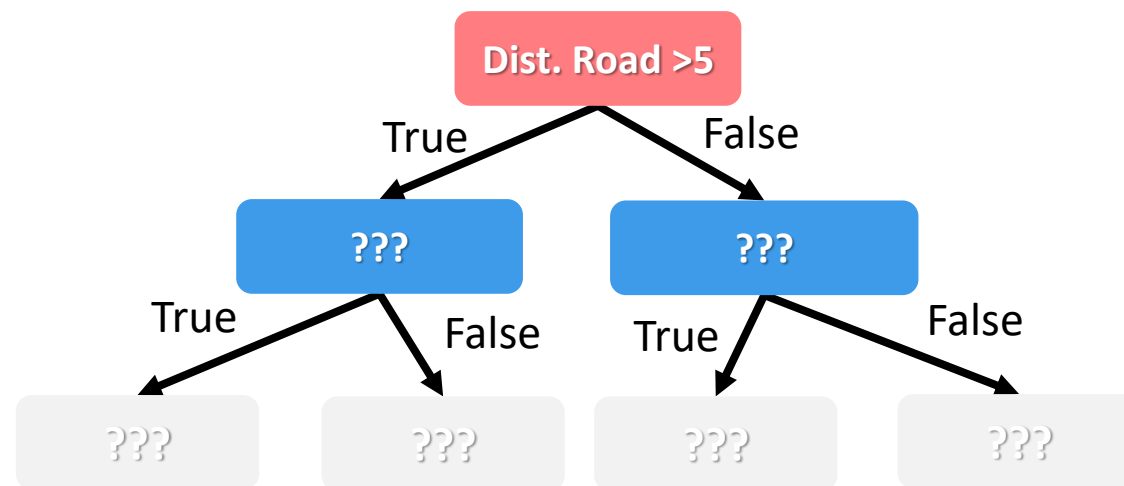
Slope <15

Good/bad soil

A melhor variável torna-se o Root Node!

Como as Árvores de Decisão são construídas?

Sample ID	Dist. City	Dist. Road	Slope	Soil qual.	Forest loss
20	18.9	4.5	12.1	Good	Yes
3	18.9	3.5	2.3	Good	No
15	1.4	3.8	2	Bad	Yes
21	6.8	2.3	13	Bad	No
50	9.8	1.1	0.4	Good	Yes
1	21.4	3.2	12	Good	Yes
34	23	5.9	3.5	Good	Yes
22	18	4.7	9.8	Bad	No
49	5.8	2.3	0.8	Bad	No
21	6.8	2.3	13	Bad	No
2	25.6	14.1	0.5	Bad	No
20	18.9	4.5	12.1	Good	Yes
7	2.6	17.1	1.6	Good	Yes
50	9.8	1.1	0.4	Good	Yes
...	...	...	...	...	...
34	23	5.9	3.5	Good	Yes
21	6.8	2.3	13	Bad	No
46	17.2	24.5	17	Good	No
4	7.8	5.6	11	Good	Yes

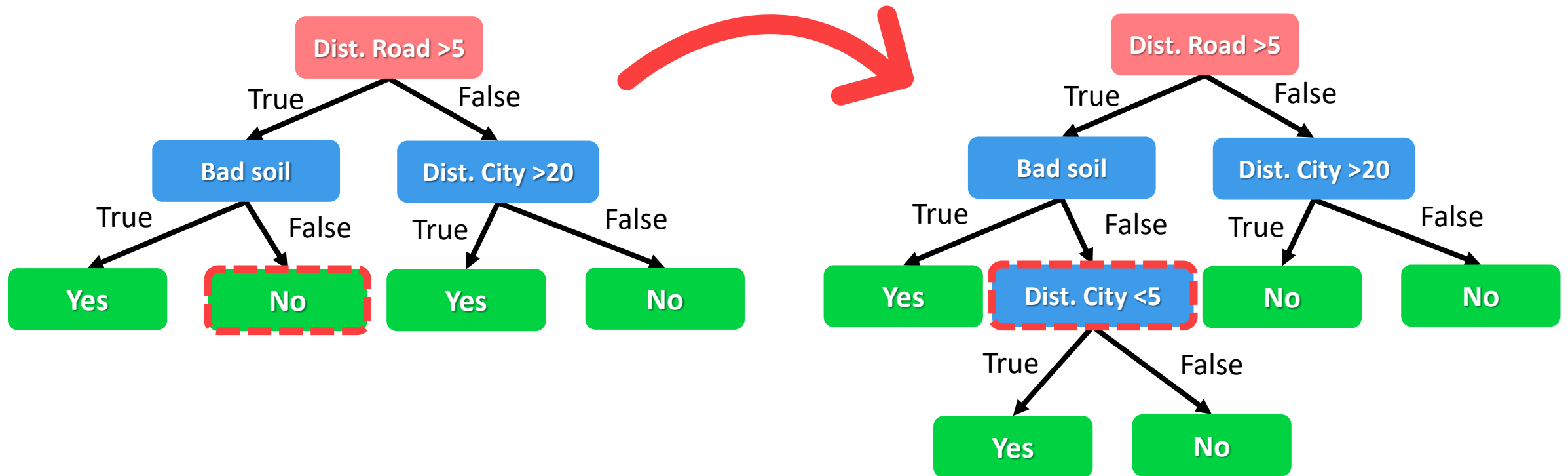


*As mesmas etapas são usadas para selecionar as variáveis (e limites) a serem usadas nos **Decision Nodes***

Obs.: “Dist. Road” com um limite diferente pode ser usado novamente no **Decision Nodes**! E lembre-se: nem todas as variáveis são consideradas em cada **Node**!!!

Como as Árvores de Decisão são construídas?

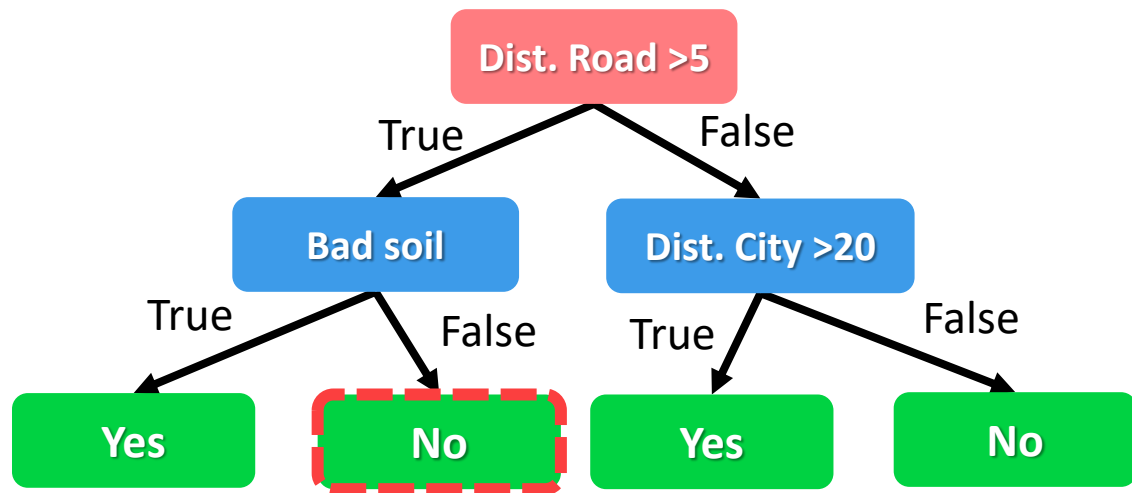
Quando o algoritmo para de dividir uma árvore (de classificação)?



Como as Árvores de Decisão são construídas?

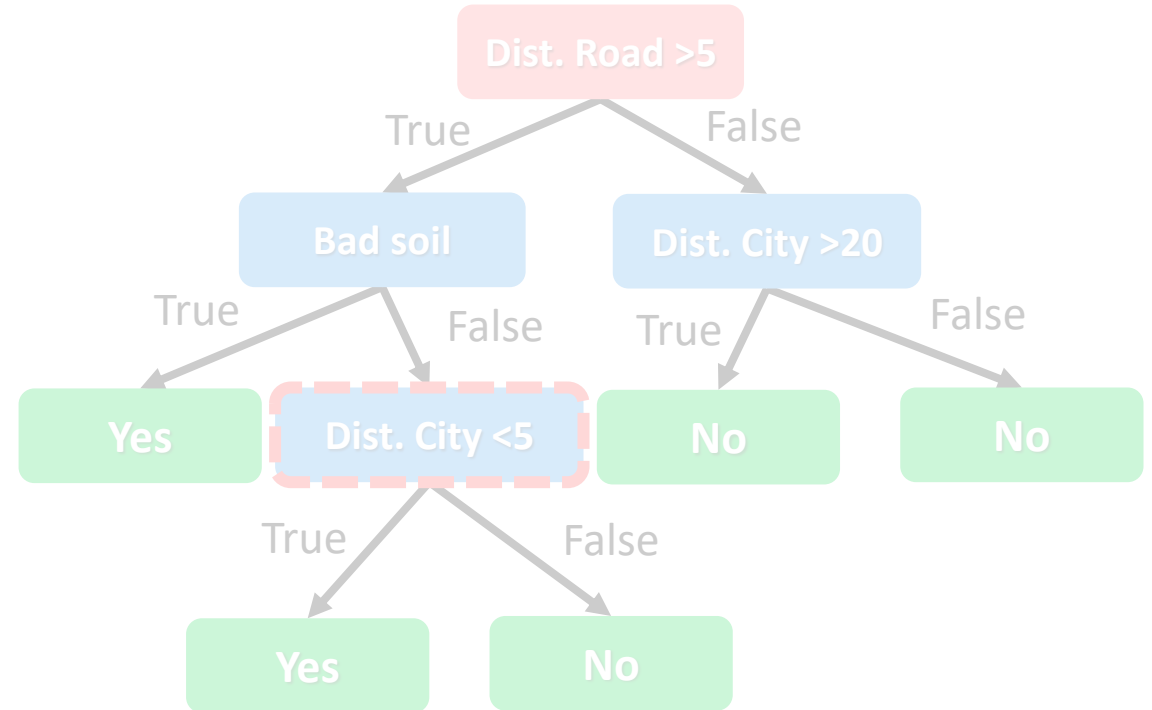
Impureza de Gini: quão bem a árvore de decisão divide os dados

(somente para árvores de classificação)



3 yes/22 no: Boa separação!

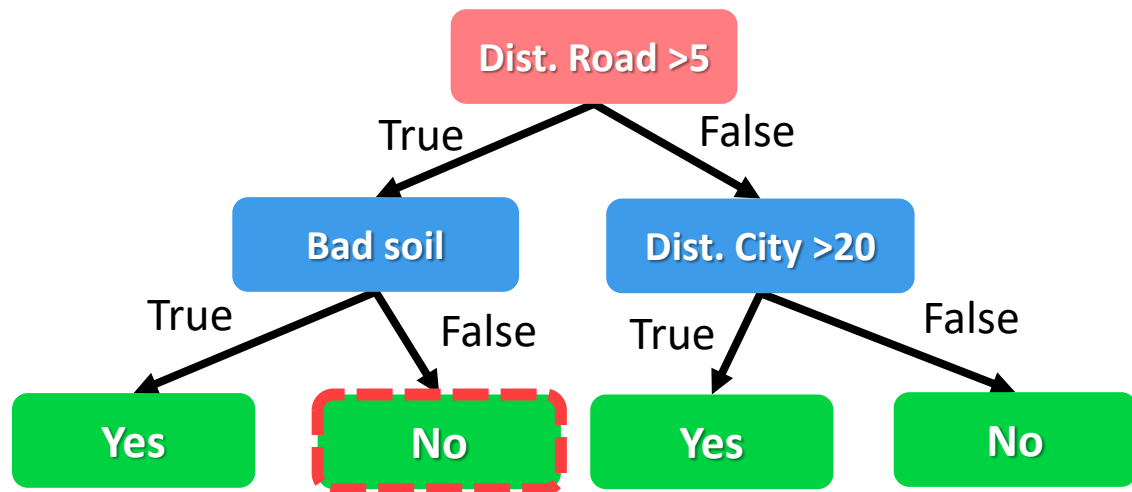
(Assim, podemos parar nesse *leaf node*)



Como as Árvores de Decisão são construídas?

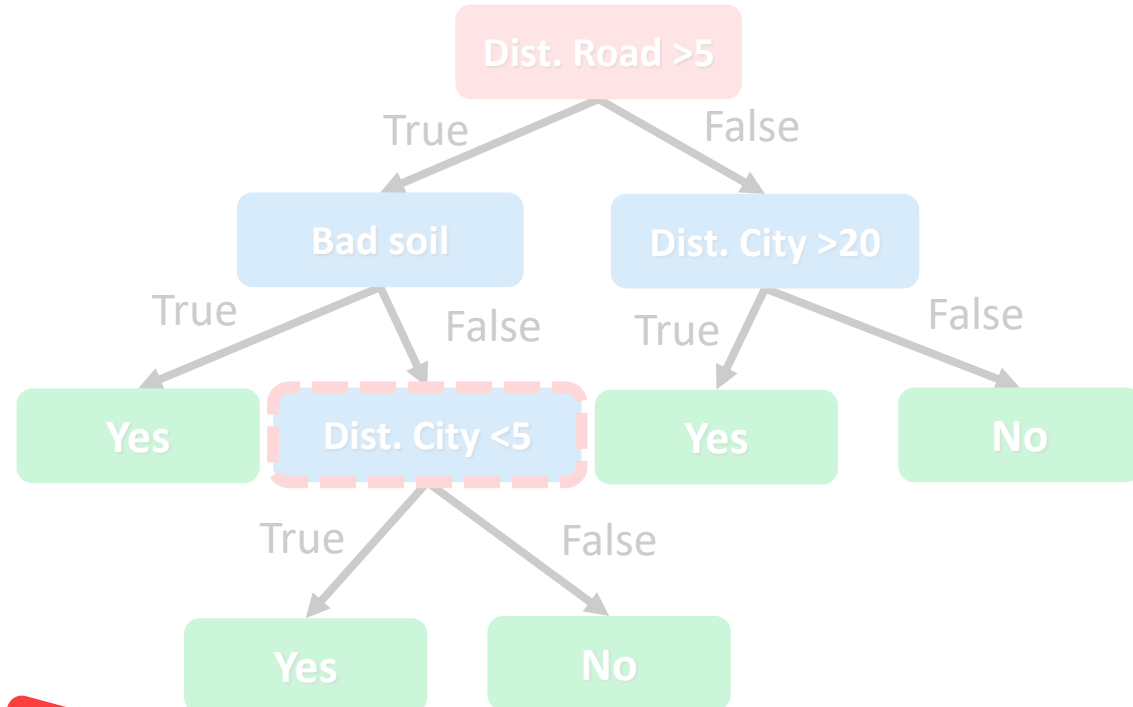
Impureza de Gini: quão bem a árvore de decisão divide os dados

(somente para árvores de classificação)



~~3 yes/22 no: Boa separação!~~

11 yes/14 no: Separação ruim...



9 yes/1 no

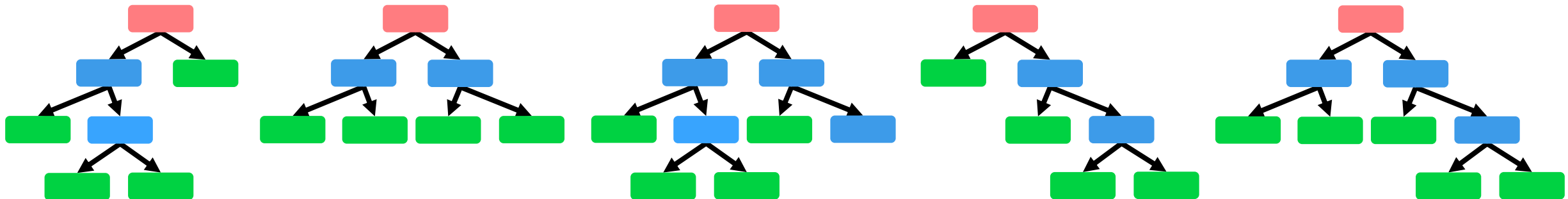
2 yes/13 no: Boa separação!

Como as Árvores de Decisão são construídas?

Hiperparâmetros também controlam a aparência das árvores finais:

- *Número mínimo de observações (dos dados de treinamento) permitido em cada **leaf node** após cada divisão*
- *“Profundidade” (“depth”) máxima das árvores*
- *Número máximo de variáveis consideradas em cada **Decision Node***

(O número de árvores de decisão da Floresta Aleatória também é um hiperparâmetro)



Agenda

- ☑ *Criando uma árvore de decisão com bootstrapping*
- ☑ *Usando um modelo de floresta aleatória*
- ☑ *Construindo um modelo de Floresta Aleatória*
- ☐ ***Florestas aleatórias de regressão***

Árvores de decisão

Umidade abaixo de 30%

Verdadeiro

Falso

135 esperados
Incêndios

85 esperados
Incêndios

*Esta árvore prevê valores
numéricos...*

É uma árvore de regressão!

Umidade abaixo de 30%

Verdadeiro

Falso

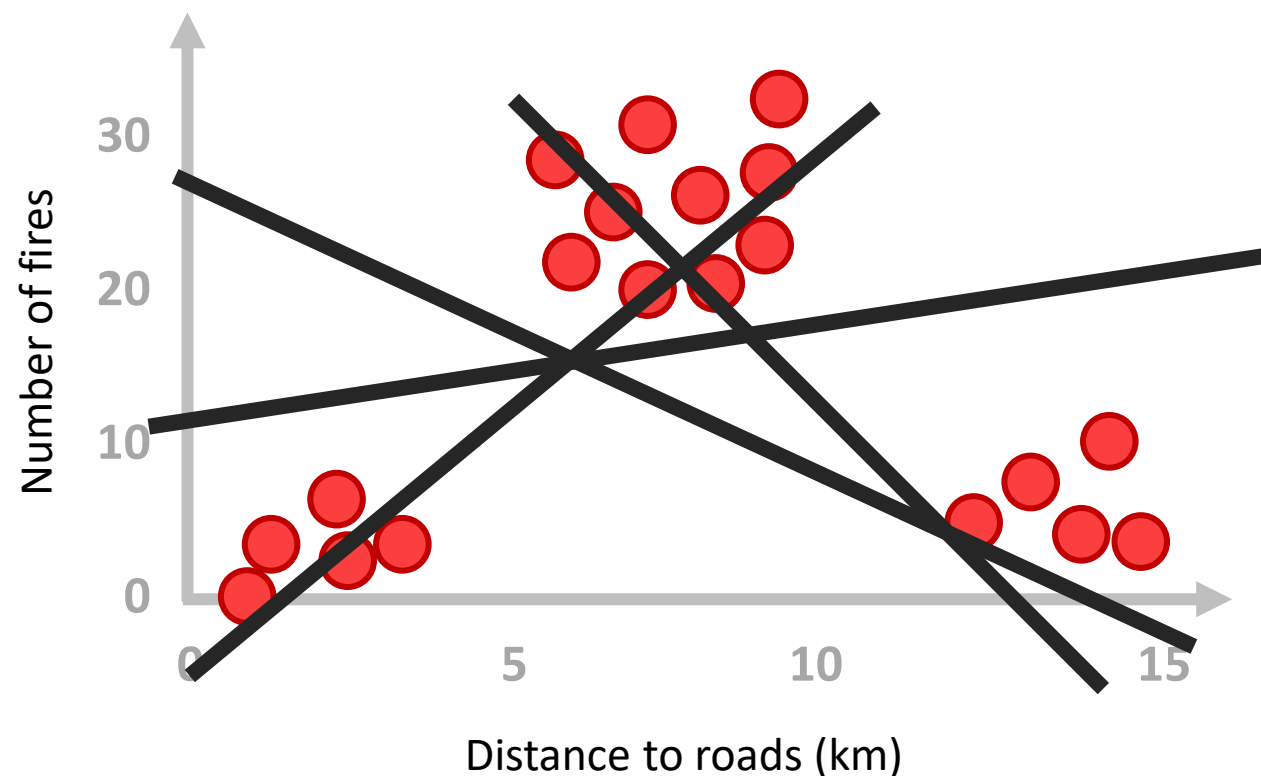
Alto risco de
Incêndios

Baixo risco de
Incêndios

*Esta árvore classifica dados
em categorias...*

É uma Árvore de Classificação!

E as árvores de regressão?

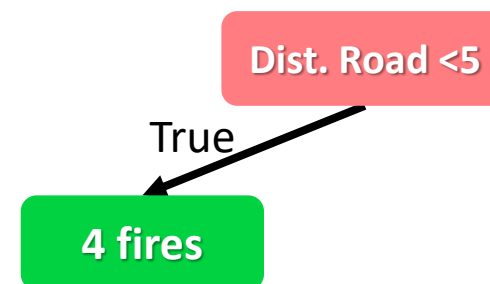
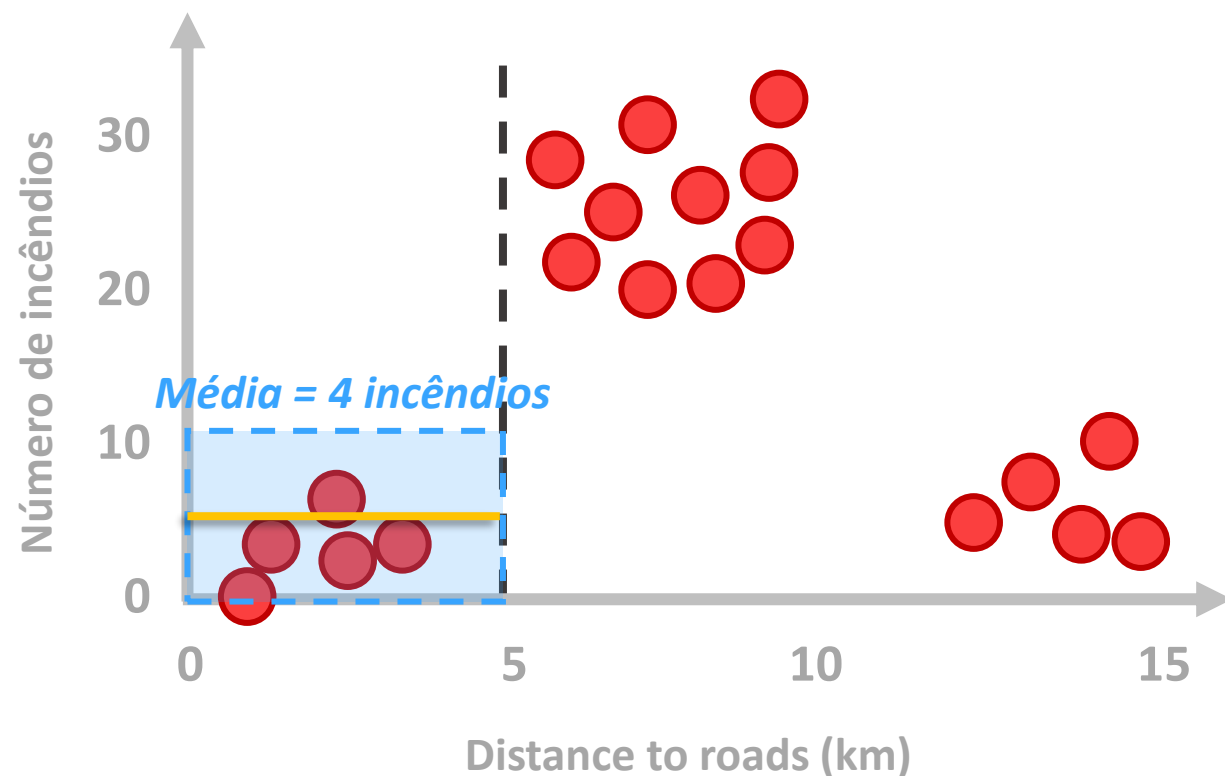


Uma linha reta (de um modelo de regressão linear) talvez não nos ajude a fazer boas previsões em alguns casos...

Mas um modelo de Floresta Aleatória baseada em árvores de regressão talvez possa ajudar!

E as árvores de regressão?

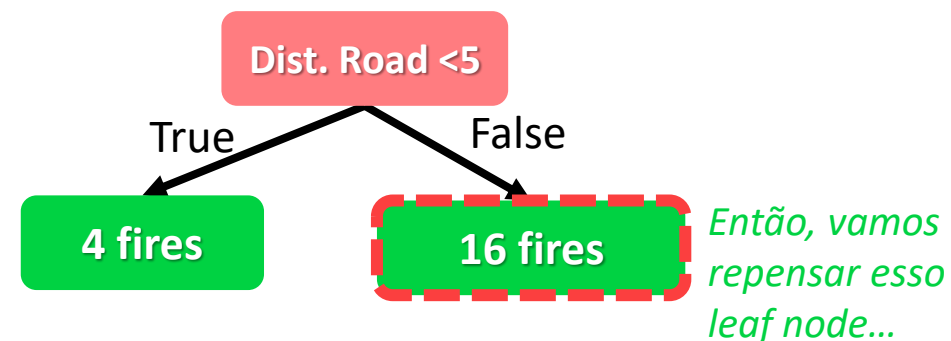
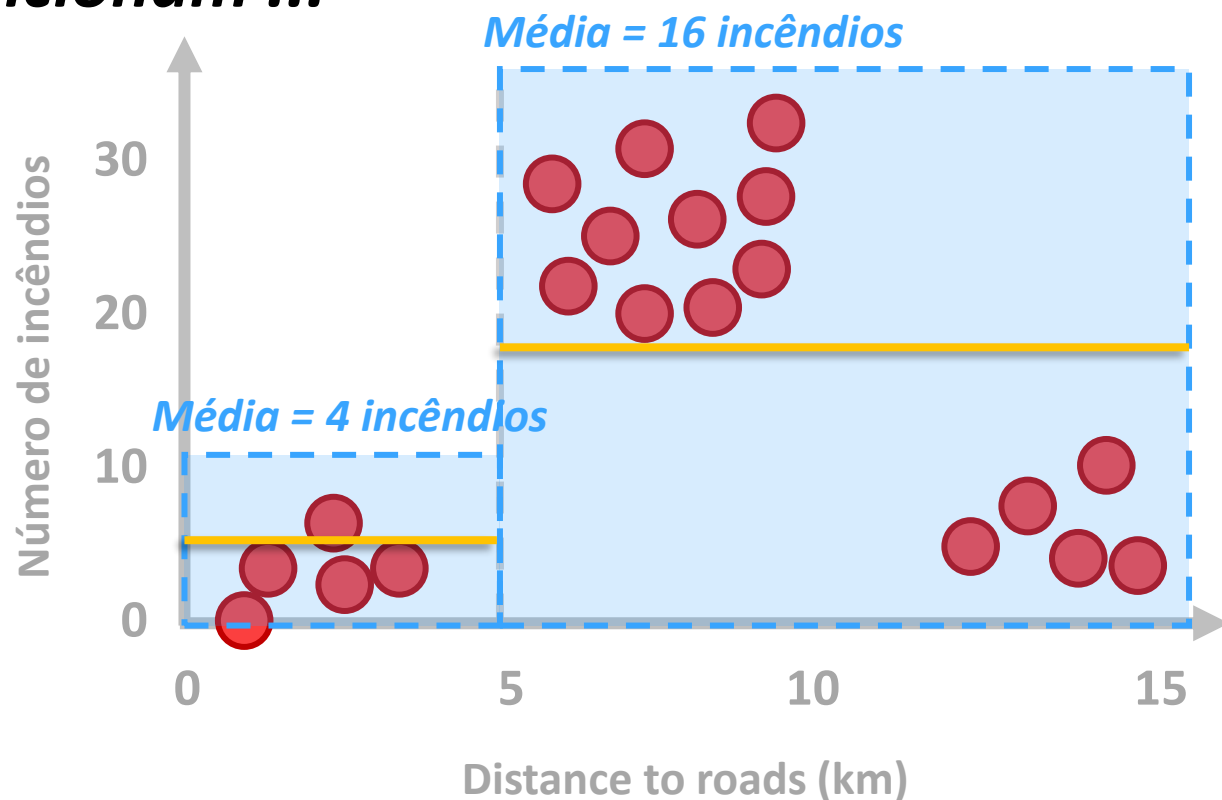
Semelhante a como as árvores de classificação funcionam ...



*Primeiro, ele identifica o limite ideal para dividir os dados, e.g.: **Dist. Road < 5***

E as árvores de regressão?

Semelhante a como as árvores de classificação funcionam ...

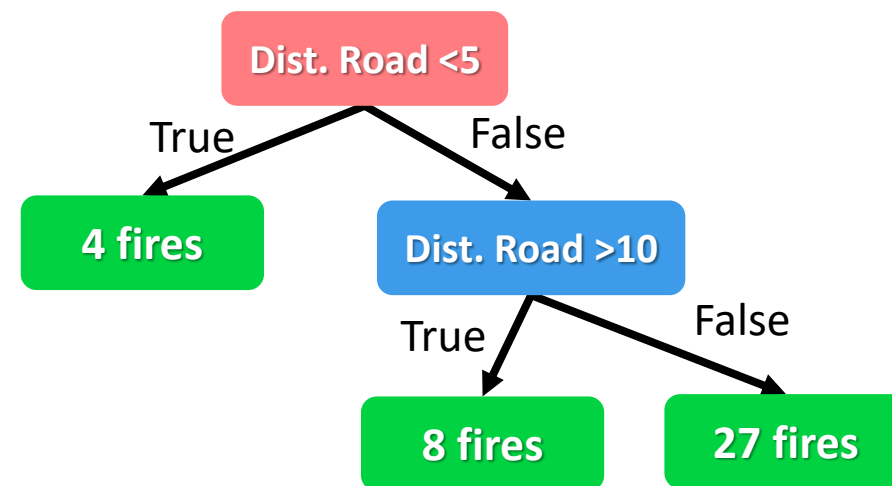
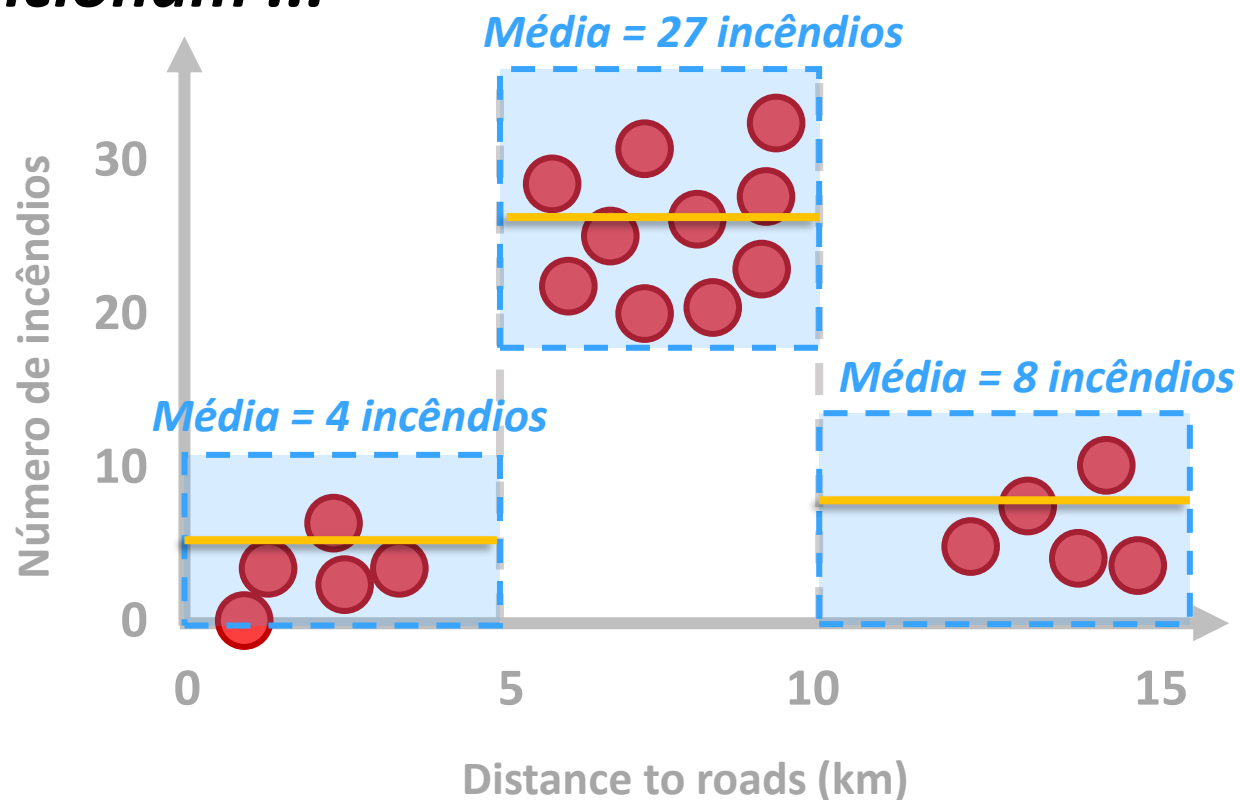


*Agora, a árvore poderia usar a mesma lógica e apenas calcular uma média em todas as amostras com **Dist. Road > 5***

*Mas há muita variação dentro do grupo "**Dist. Road > 5**"... Podemos fazer melhor do que isso, certo?*

E as árvores de regressão?

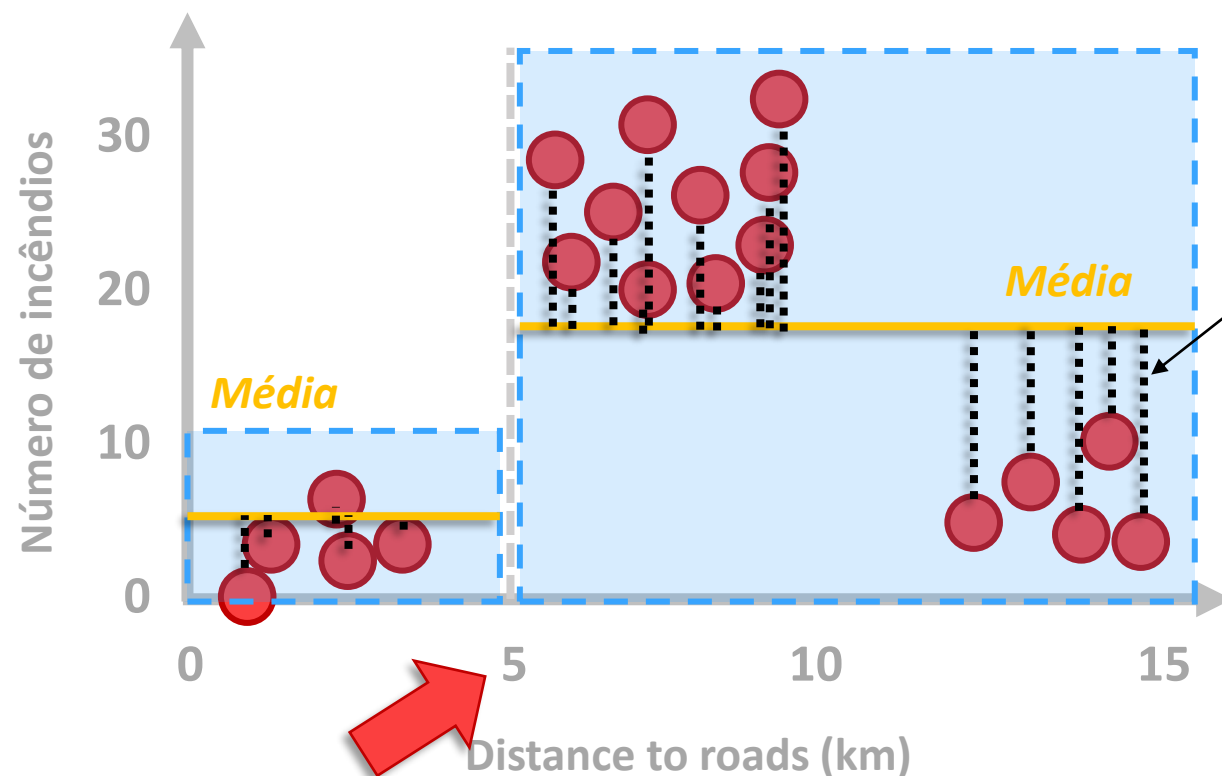
Semelhante a como as árvores de classificação funcionam ...



*Em vez disso, vamos identificar outros limites para dividir ainda mais os dados (quantas vezes forem necessárias) e.g.: **Dist. Road > 10***

E as árvores de regressão?

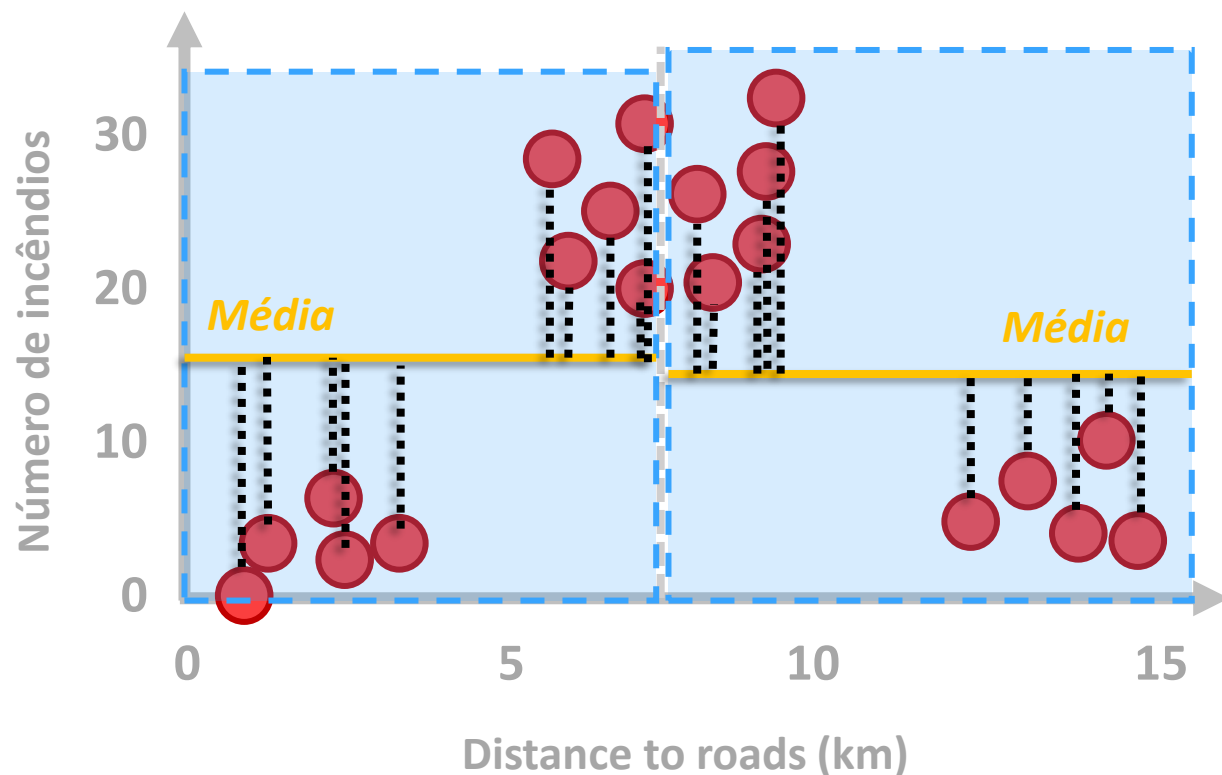
Como os limites ótimos são identificados?



Estes são os *resíduos*!
Normalmente, usamos a **Soma dos Resíduos Quadrados (SRQ)** associada a diferentes limiares e escolhemos o limite com o **SRQ** mais baixo (= limite ideal)

What about Regression Trees?

Como os limites ótimos são identificados?



~~Dist. Road > 7~~

*Por exemplo, se usarmos 7 em vez de 5 como limite, teremos novas **médias** para os grupo, que estão associadas a um **SRQ** mais alto...*

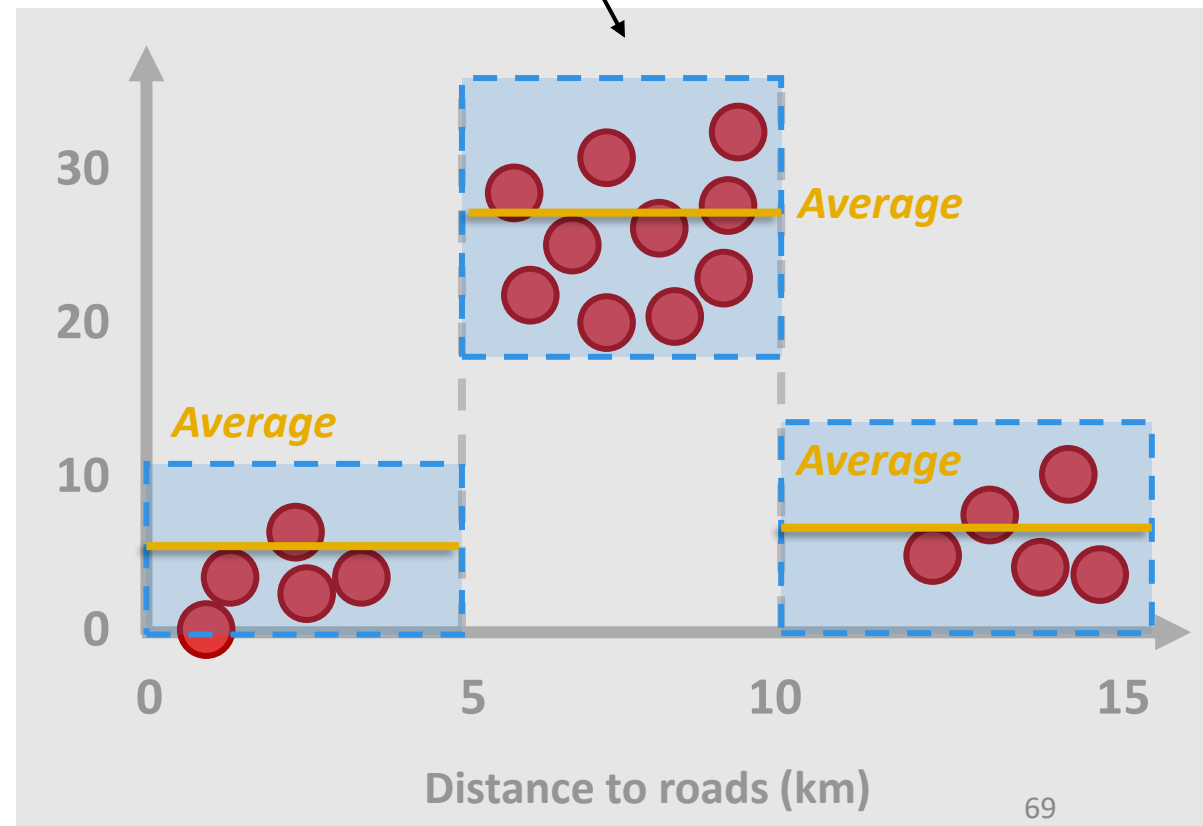
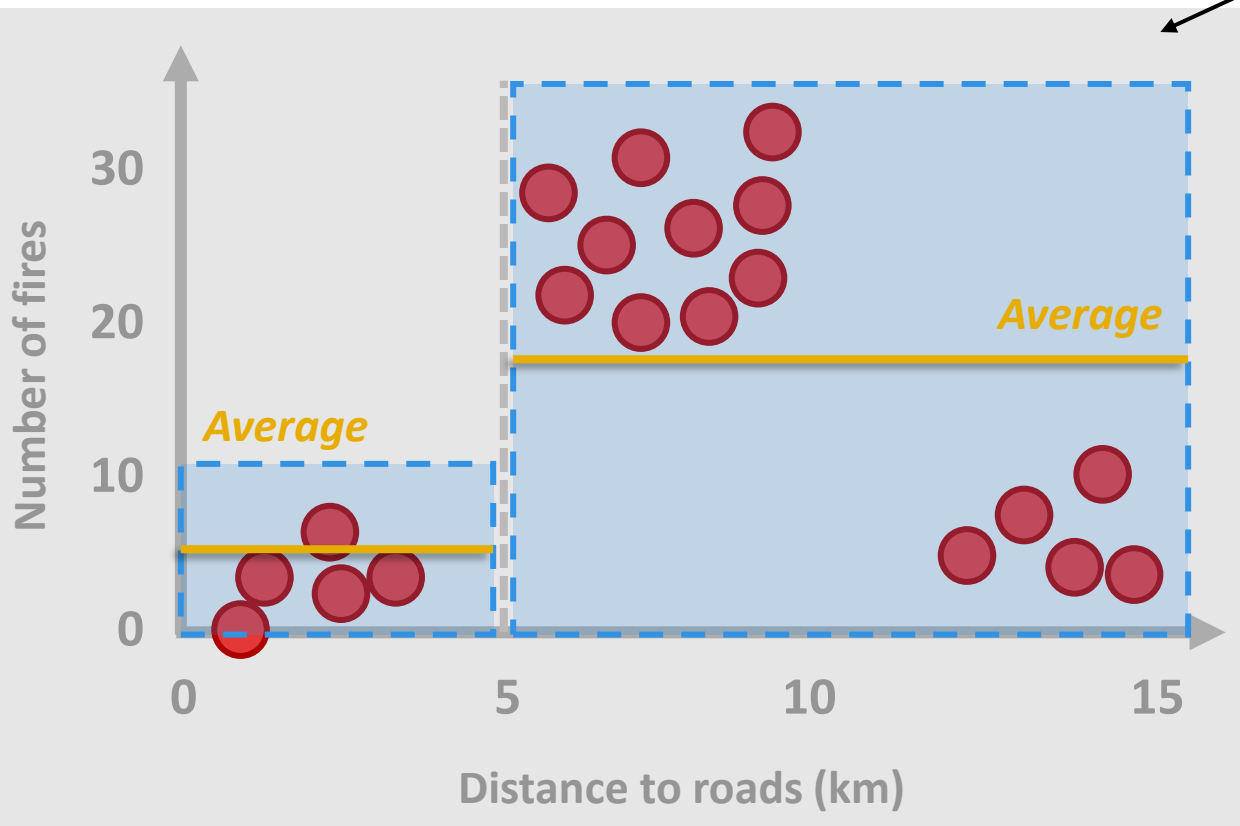
Portanto, 7 não é o limite ideal!

Obs.: alguns algoritmos podem usar uma métrica diferente do SRQ

E as árvores de regressão?

E quando parar de dividir na árvore?

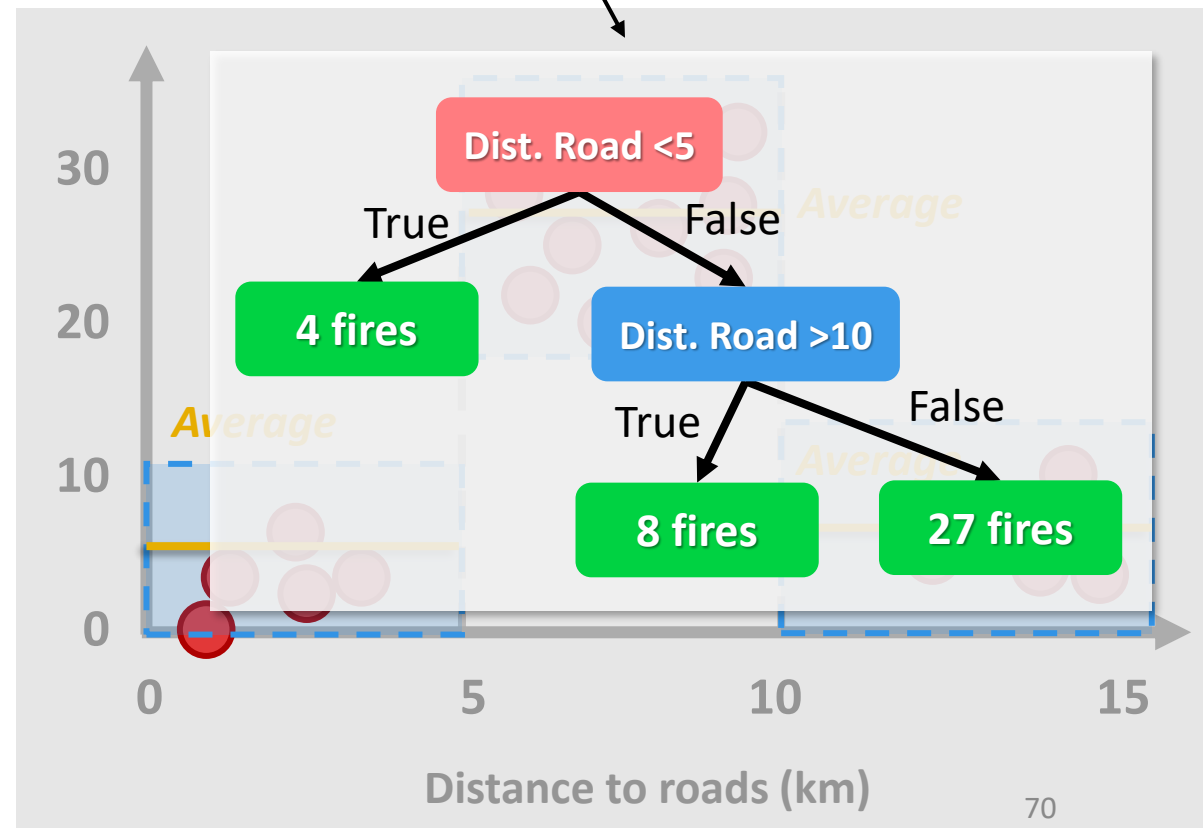
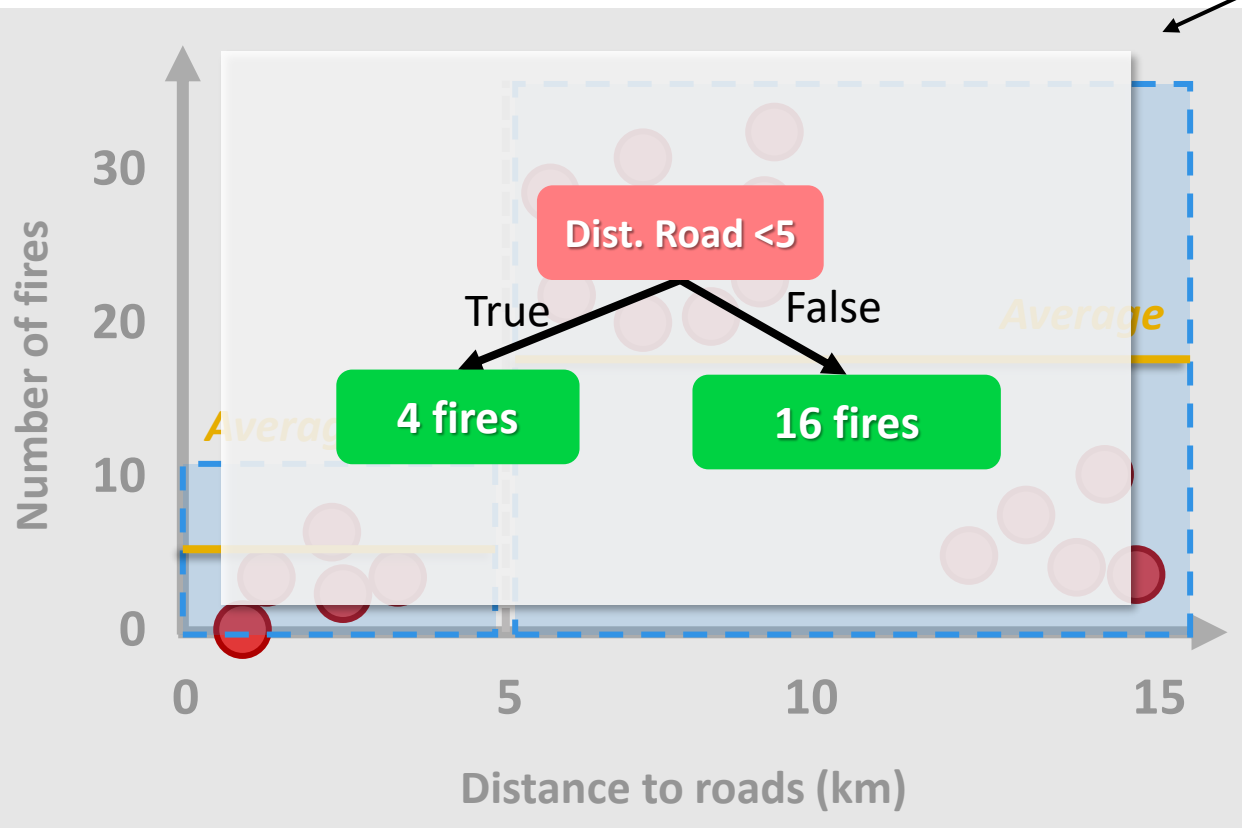
*Temos um **SRQ** mais baixo com 2 grupos ou 3 grupos?*



E as árvores de regressão?

E quando parar de dividir na árvore?

*Temos um **SRQ** mais baixo com 2 grupos ou 3 grupos?*



E as árvores de regressão?

E quando parar de dividir na árvore?

Para as árvores de regressão, dividir os dados em mais grupos sempre reduzirá o SRQ... Portanto, são os hiperparâmetros que determinam quando as árvores de regressão param de se dividir

*Sem os hiperparâmetros, as árvores de regressão continuariam se dividindo até criarem **leaf nodes** contendo apenas uma única observação. Esse cenário resultaria em um modelo altamente complexo e provavelmente sobreajustado (“overfitted”)*

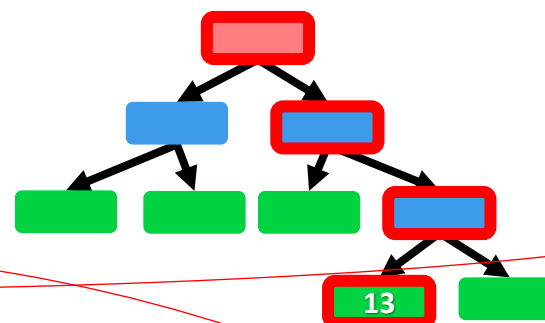
Como usamos uma floresta aleatória (de regressão)?

Semelhante ao modelo de Floresta Aleatória de classificação :

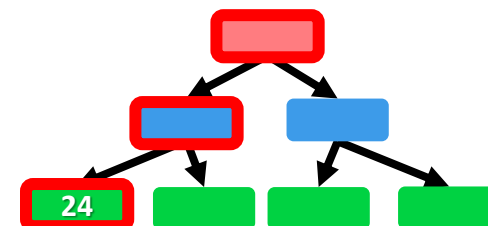
Sample ID	Variable 1	Variable 2	Variable 3	Variable 4	# fires
"New"	8.9	14	22.1	Good	13

Quantos incêndios previstos?
É a média de 13, 24, 8, 12, 15,
11 = 13

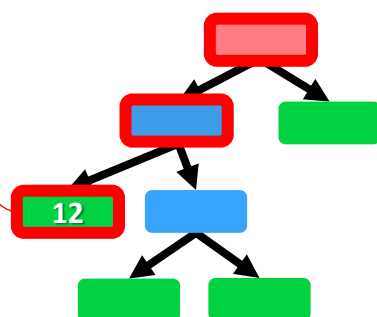
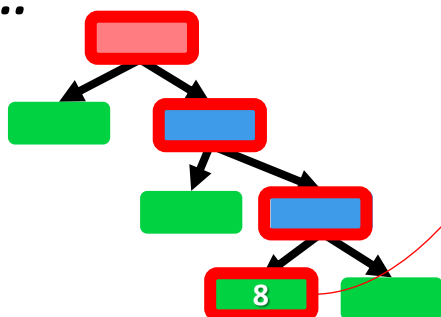
Passamos os dados novos pela primeira árvore...



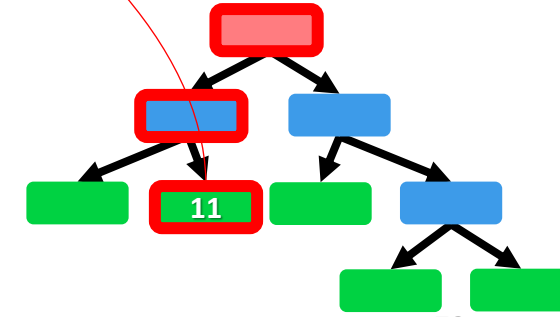
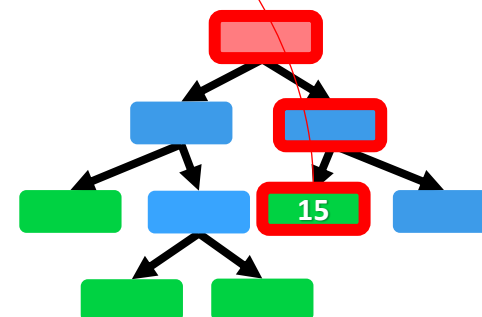
... e através da segunda árvore...



... e através da terceira árvore...



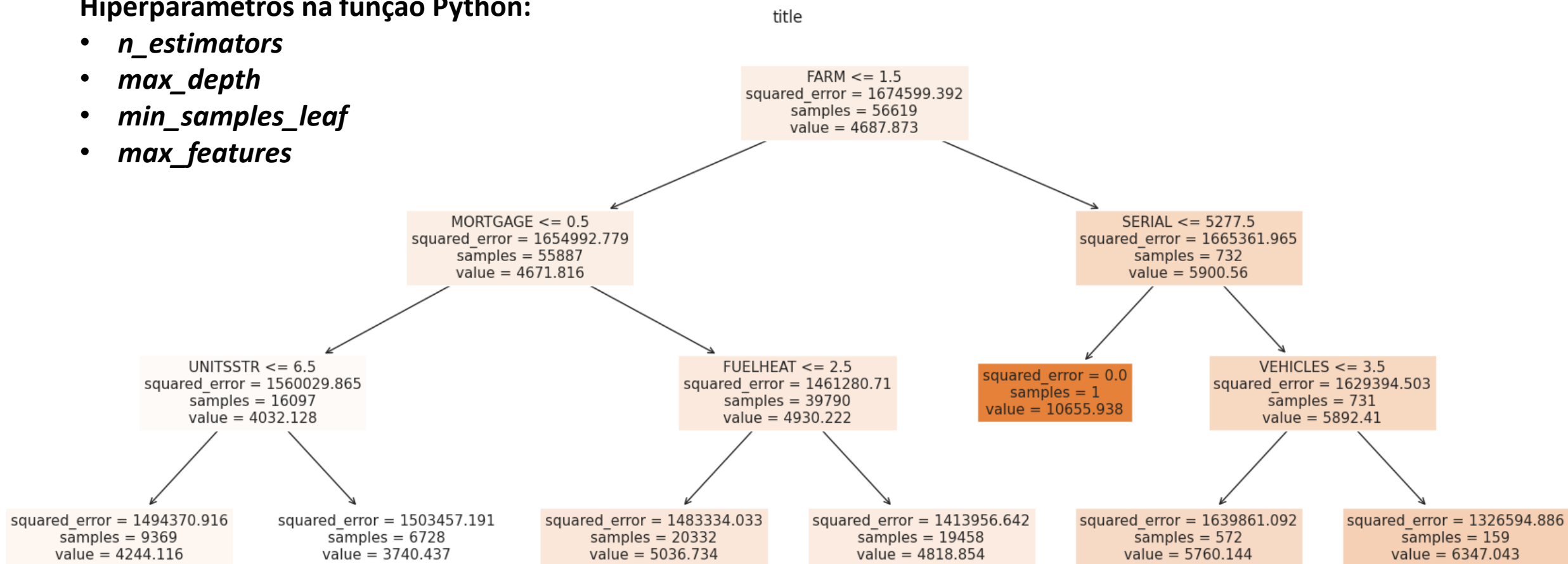
... e assim por diante... E calculamos uma média



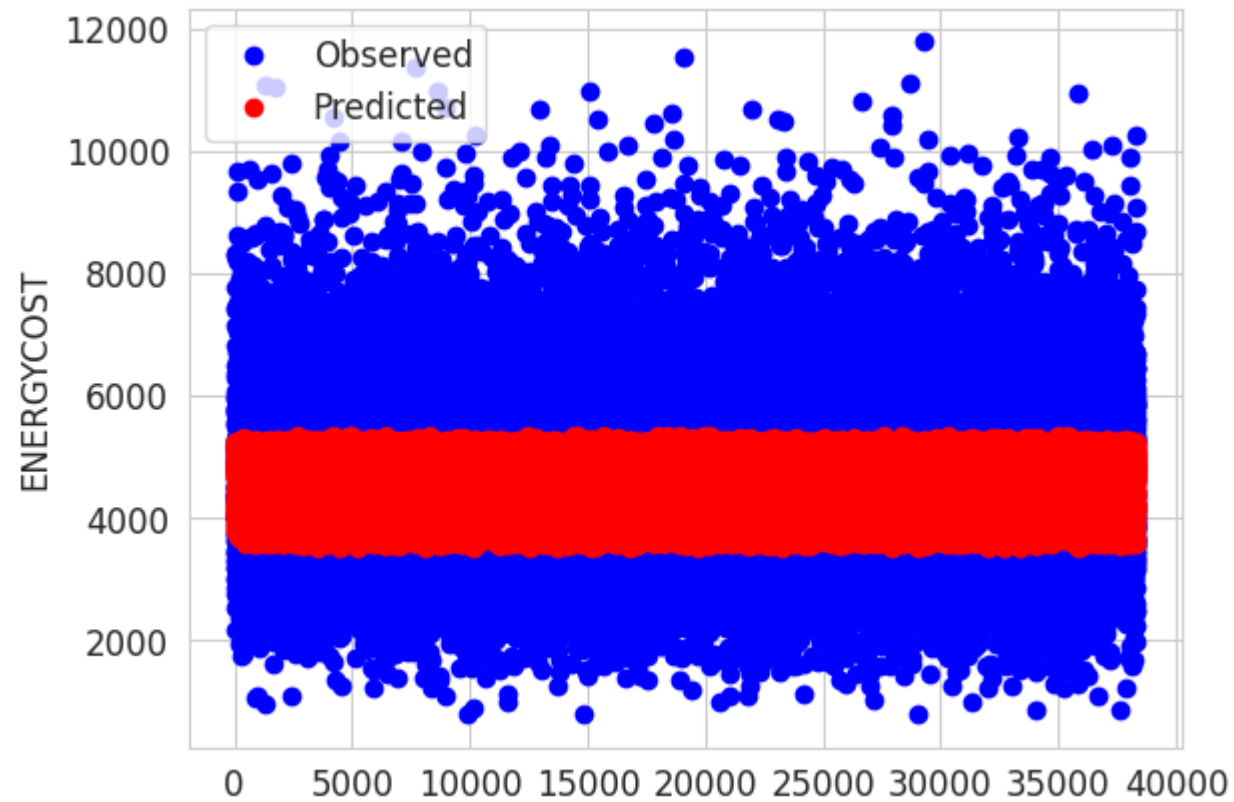
Exemplo de uma árvore criada em Python

Hiperparâmetros na função Python:

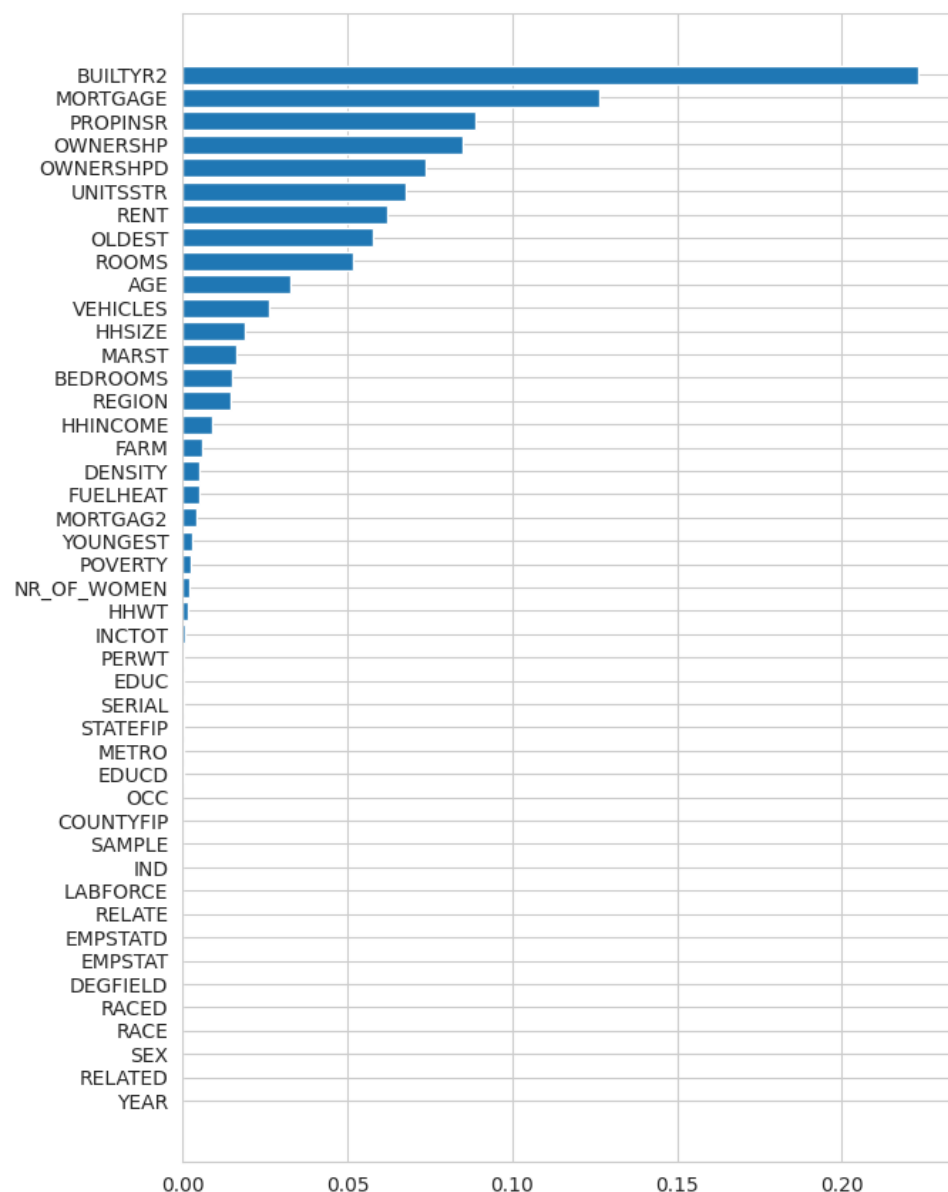
- *n_estimators*
- *max_depth*
- *min_samples_leaf*
- *max_features*



Exemplo: valores observados vs. simulados



Exemplo: seleção de “features”



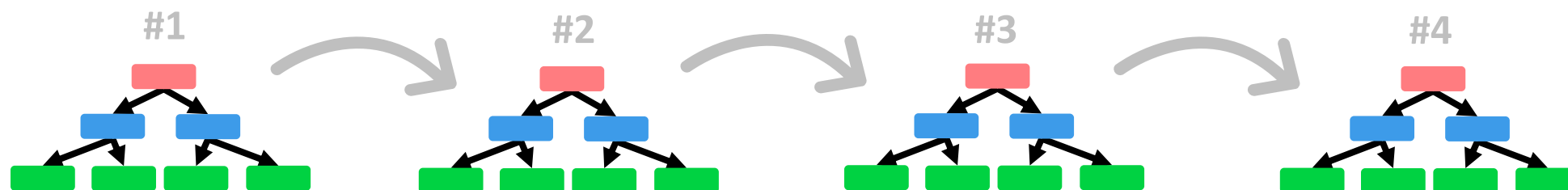
*Depois de treinar a floresta aleatória, normalmente você pode obter pontuações de importância das **features (variáveis)** do modelo*

Essas pontuações refletem a diminuição média na impureza para as árvores de classificação (por exemplo, Impureza de Gini) ou resíduos para as árvores de regressão (por exemplo, SRQ) em todas as árvores no modelo de floresta aleatória quando um recurso específico é usado para divisão

*Você pode usar essas pontuações para identificar as **features (variáveis)** mais importantes em seu modelo*

Extra: Floresta aleatórias vs. Boosting Decision Trees

Característica	Random Forest	Boosting Decision Trees
Construção das Árvores	<u>Independentes</u> , paralelas	<u>Sequenciais</u>
Árvores Individuais	<u>Profundas</u> (propensas a overfitting)	<u>Rasas</u> (“stumps” de decisão)
Amostragem de Dados	Amostragem <u>Bootstrap</u> (com reposição)	Todos os dados, com pesos ajustados com base nos erros
Amostragem de Features	<u>Subconjunto</u> aleatório de features a cada divisão	Tipicamente <u>todas</u> as features a cada divisão de um aprendiz fraco
Combinação	<u>Média ou voto majoritário</u> ao final	<u>Soma ponderada</u> das previsões sequencialmente
Correção de Erros	Reduz a variância através da média	Reduz o viés corrigindo erros anteriores
Robustez a Ruídos	Geralmente mais robusto	Pode ser mais sensível
Interpretabilidade	Menor que uma única árvore de decisão	Geralmente menor que uma Random Forest
Treinamento	Paralelizável	Sequencial



Resumo

- *As florestas aleatórias são baseadas em uma coleção de árvores de **classificação** ou **regressão***
- *Eles são baseados em **bootstrapping + agregação ("bagging")**, que é um tipo de modelo de conjunto ("assemble"), ou seja, um modelo que combina as previsões de vários modelos individuais, neste caso, árvores individuais, o que ajuda a evitar o sobreajuste ("overfitting")*
- *Os **hiperparâmetros** influenciam diretamente como as árvores são construídas*
- *As técnicas de **seleção de features** podem ser usadas para identificar os recursos mais importantes (ou seja, variáveis) em um modelo de floresta aleatória*