Faculdade XPe

Fundamentos de Machine Learning

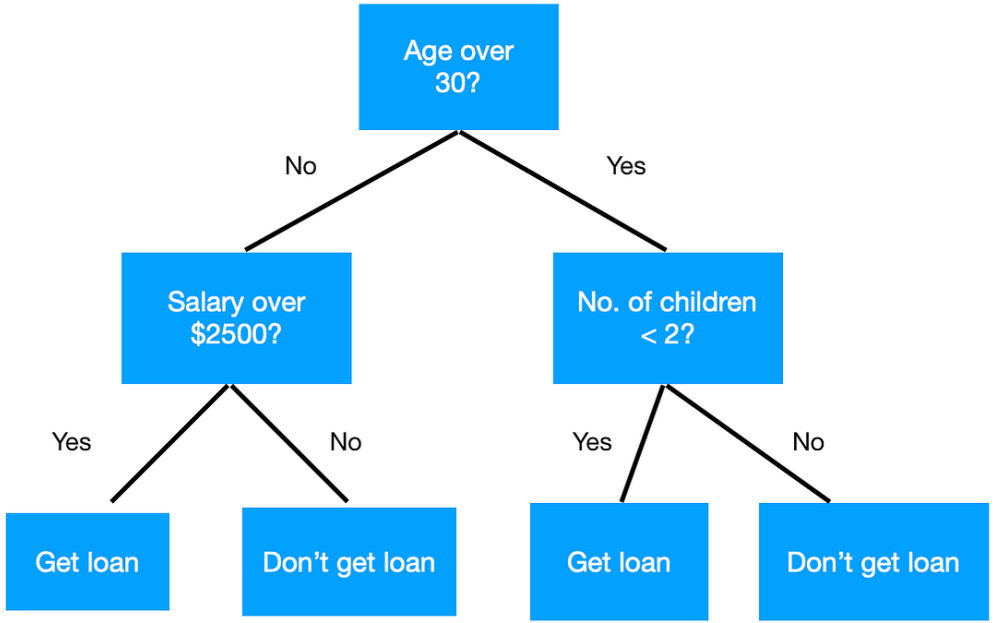
# Comitês de Modelos

Aula 1.1  Medindo erros

Prof. Pedro Calais

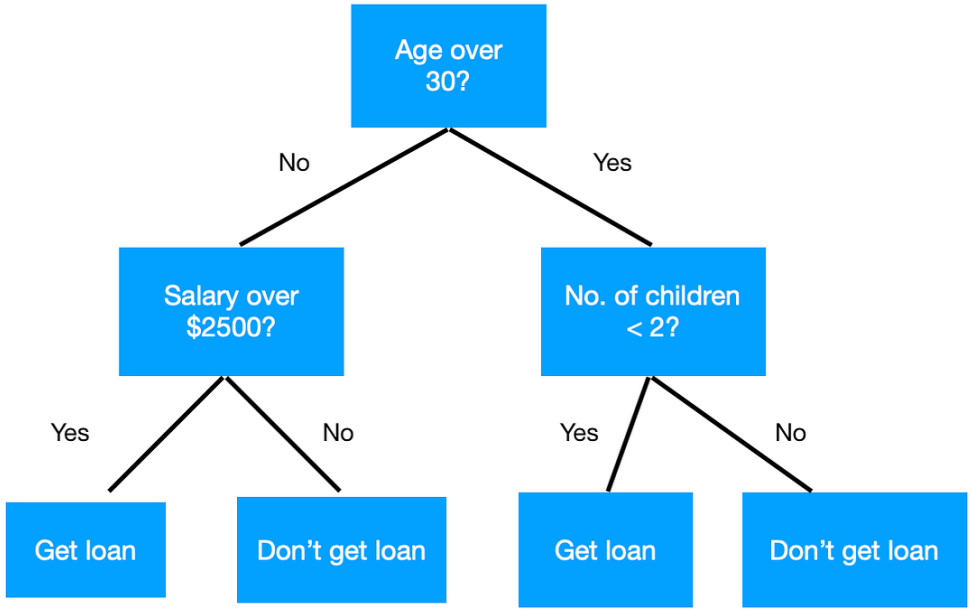
Faculdade XPe

# Lembra da árvore de decisão?



Faculdade XPe

- Fácil de construir
- Fácil de usar
- Fácil de interpretar



Faculdade XPe

# Problema: árvores de decisão são instáveis

## Instability of Decision Tree Classification Algorithms

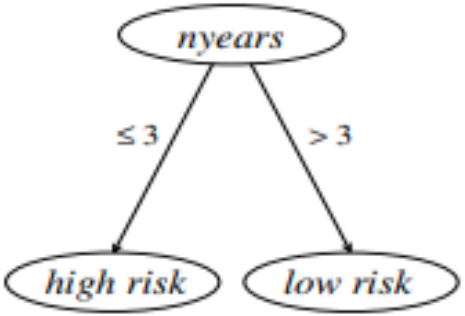
Ruey-Hsia Li
Lightspeed Semiconductor
209 N. Fair Oaks Avenue
Sunnyvale, CA 94085
rli@lightspeed.com

Geneva G. Belford
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
belford@cs.uiuc.edu

Faculdade XPe

# Um pequeno conjunto de dados

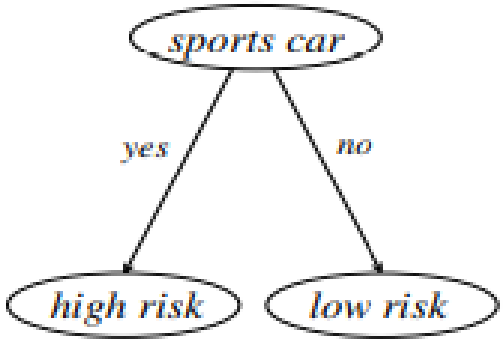
| nyears | sports car | risk |
|---|---|---|
| 1 | *yes* | *high* |
| 2 | *no* | *high* |
| 2 | *yes* | *high* |
| 4 | *no* | *low* |
| 8 | *no* | *low* |



Faculdade XPe

# E se o registro #2 for diferente?

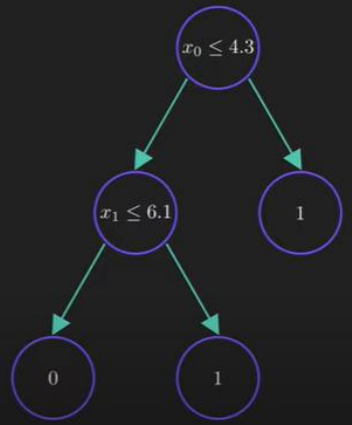
| nyears | sports car | risk |
|---|---|---|
| 1 | *yes* | *high* |
| 4 | *yes* | *high* |
| 2 | *yes* | *high* |
| 4 | *no* | *low* |
| 8 | *no* | *low* |



Faculdade XPe

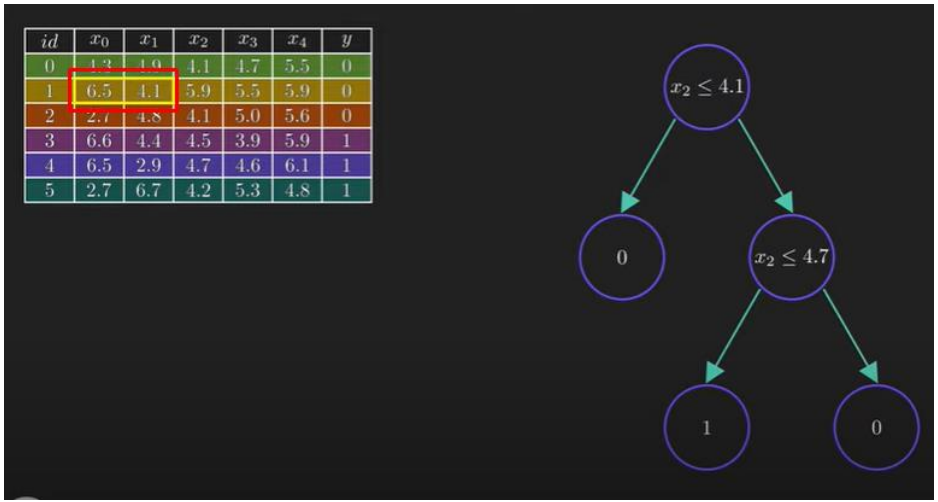
# Mais um exemplo

| $id$ | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|---|---|
| 0 | 4.3 | 4.9 | 4.1 | 4.7 | 5.5 | 0 |
| 1 | 3.9 | 6.1 | 5.9 | 5.5 | 5.9 | 0 |
| 2 | 2.7 | 4.8 | 4.1 | 5.0 | 5.6 | 0 |
| 3 | 6.6 | 4.4 | 4.5 | 3.9 | 5.9 | 1 |
| 4 | 6.5 | 2.9 | 4.7 | 4.6 | 6.1 | 1 |
| 5 | 2.7 | 6.7 | 4.2 | 5.3 | 4.8 | 1 |




Faculdade XPe

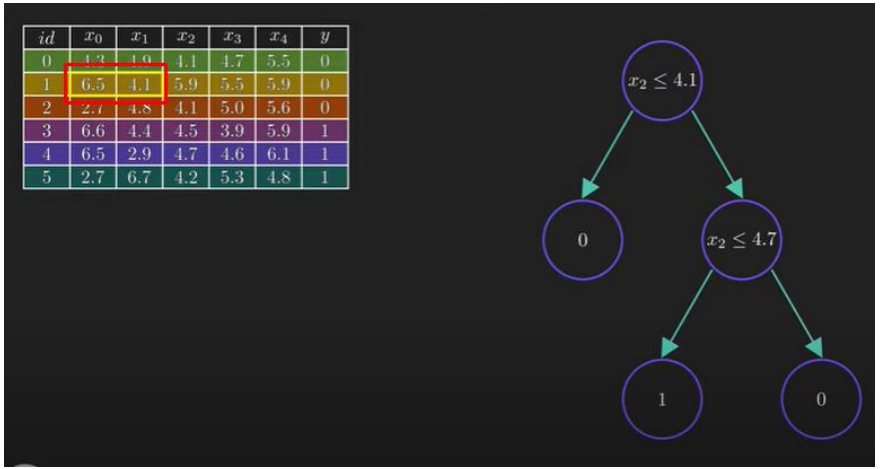
# Introduzindo uma pequena mudança



Faculdade XPe

# Árvores de decisão têm alta variância

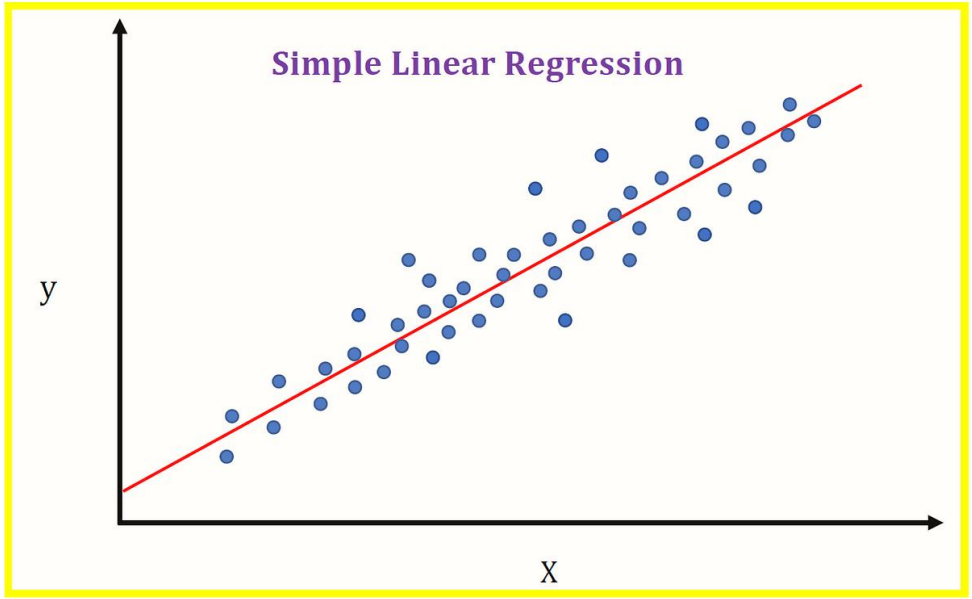
- Pequenas mudanças nos dados ou nos hiperparâmetros
- → Modelos muito diferentes



Faculdade XPe

# Contraponto

- Regressão linear tende a ter baixa variância

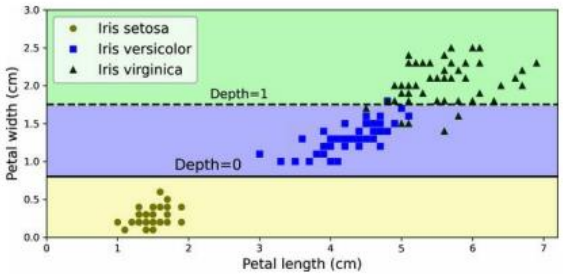


Faculdade XPe

# Árvores de decisão têm alta variância


**Decision Trees Have a High Variance**

More generally, the main issue with decision trees is that they have quite a high variance: small changes to the hyperparameters or to the data may produce very different models. In fact, since the training algorithm used by Scikit-Learn is stochastic—it randomly selects the set of features to evaluate at each node—even retraining the same decision tree on the exact same data may produce a very different model, such as the one represented in Figure 6-9 (unless you set the random_state hyperparameter). As you can see, it looks very different from the previous decision tree (Figure 6-2).



Faculdade XPe

# Por que alta variância é um problema?

- Imprevisibilidade
- Aprendendo o ruído
- Treina o modelo várias vezes, resultados muito diferentes
- Difícil validar
- Modelo está se ajustando em excesso?

Faculdade XPe

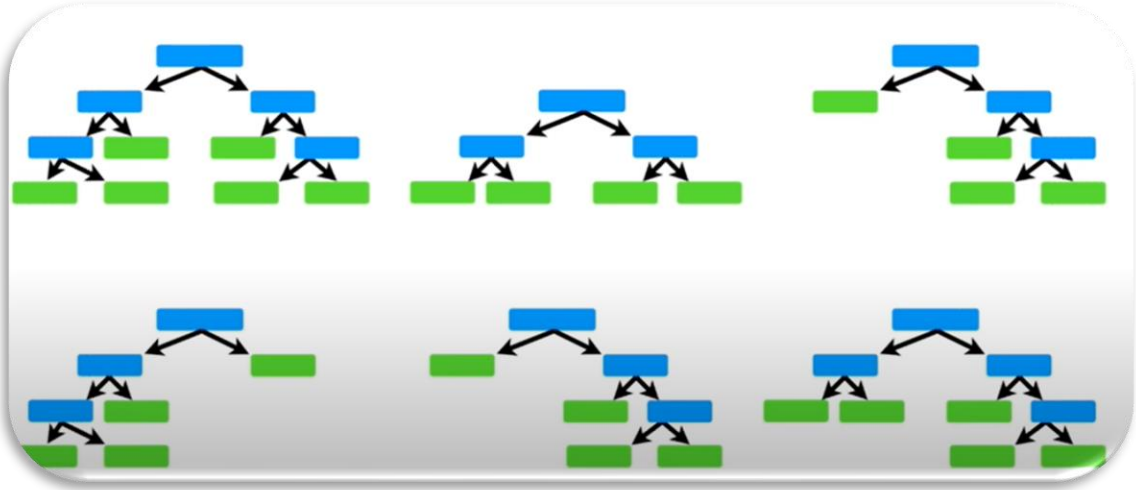
# Como reduzir a variância de árvores de decisão?

Fundamentos de Machine Learning

Faculdade XPe

# Como reduzir a variância de árvores de decisão?

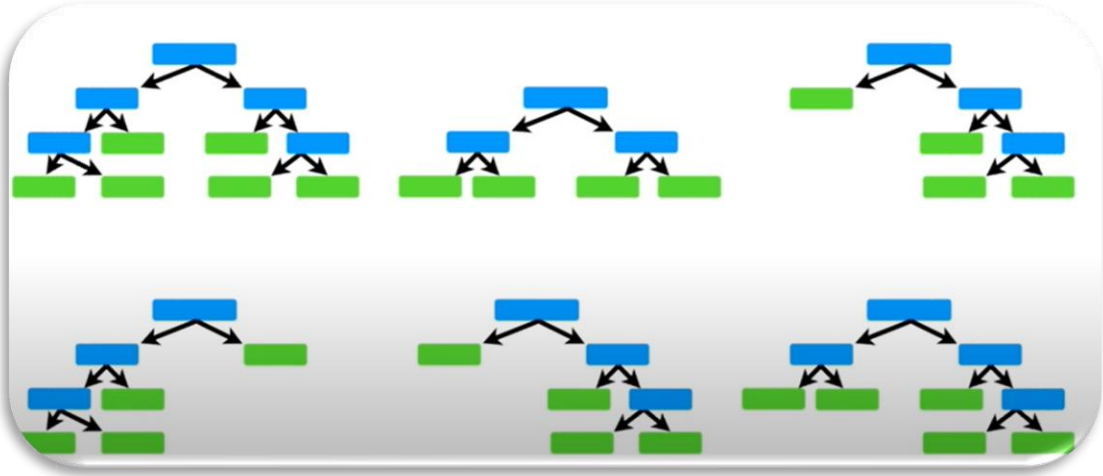
- Gere várias árvores
- Tire a média das previsões



Faculdade XPe

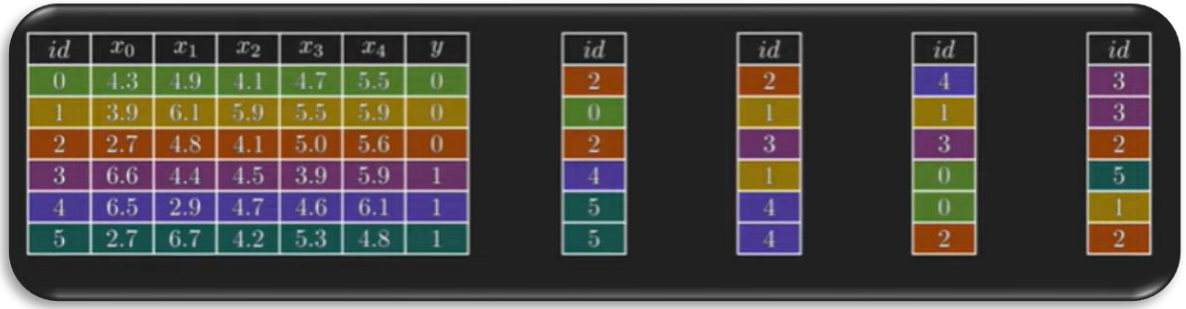
# As árvores precisam ser diferentes, certo?

- Selecione os registros aleatoriamente
- Selecione colunas aleatoriamente



Faculdade XPe

# Aleatorize os registros



| $id$ | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|---|---|
| 0 | 4.3 | 4.9 | 4.1 | 4.7 | 5.5 | 0 |
| 1 | 3.9 | 6.1 | 5.9 | 5.5 | 5.9 | 0 |
| 2 | 2.7 | 4.8 | 4.1 | 5.0 | 5.6 | 0 |
| 3 | 6.6 | 4.4 | 4.5 | 3.9 | 5.9 | 1 |
| 4 | 6.5 | 2.9 | 4.7 | 4.6 | 6.1 | 1 |
| 5 | 2.7 | 6.7 | 4.2 | 5.3 | 4.8 | 1 |

| $id$ |
|---|
| 2 |
| 0 |
| 2 |
| 4 |
| 5 |
| 5 |

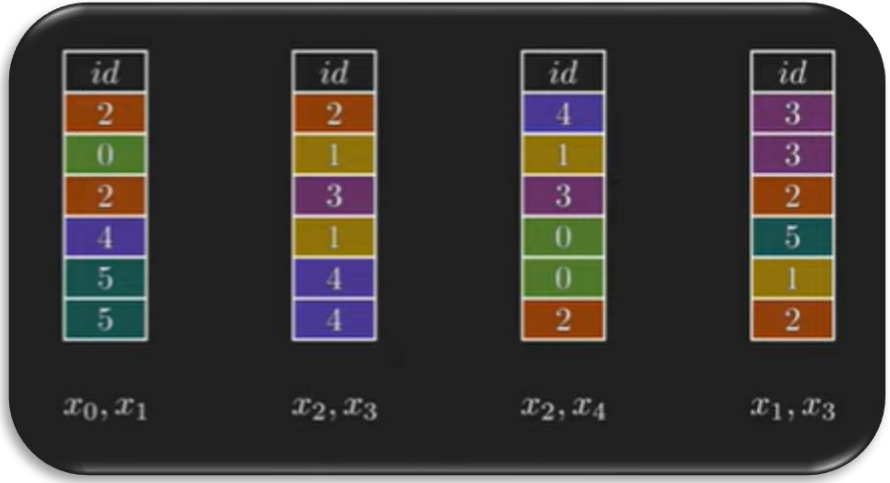
| $id$ |
|---|
| 2 |
| 1 |
| 3 |
| 1 |
| 4 |
| 4 |

| $id$ |
|---|
| 4 |
| 1 |
| 3 |
| 0 |
| 0 |
| 2 |

| $id$ |
|---|
| 3 |
| 3 |
| 2 |
| 5 |
| 1 |
| 2 |

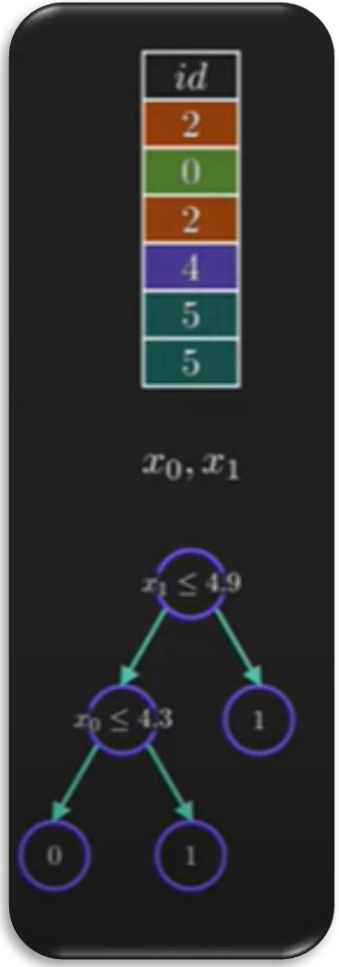
Faculdade XPe

# Aleatorize os atributos



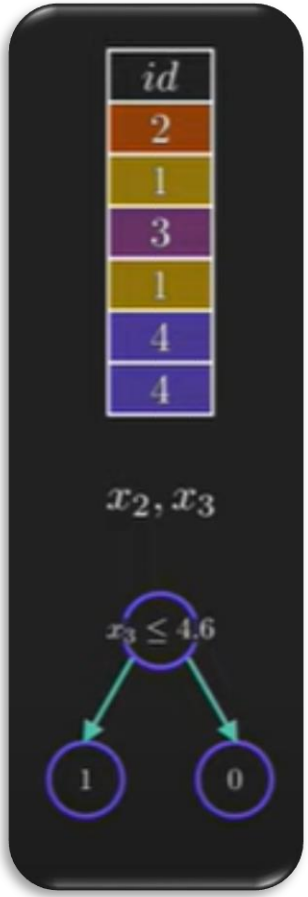
Faculdade XPe

# Árvore # 1



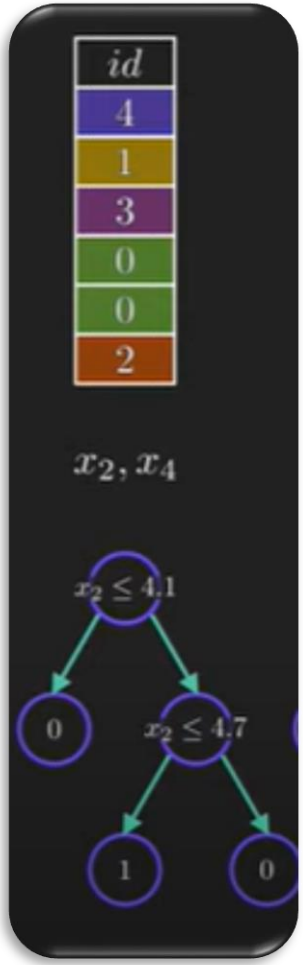
Faculdade XPe

# Árvore # 2



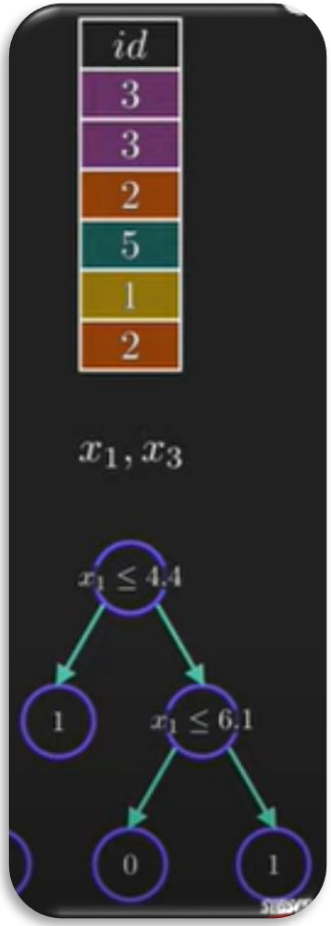
Faculdade XPe

# Árvore # 3



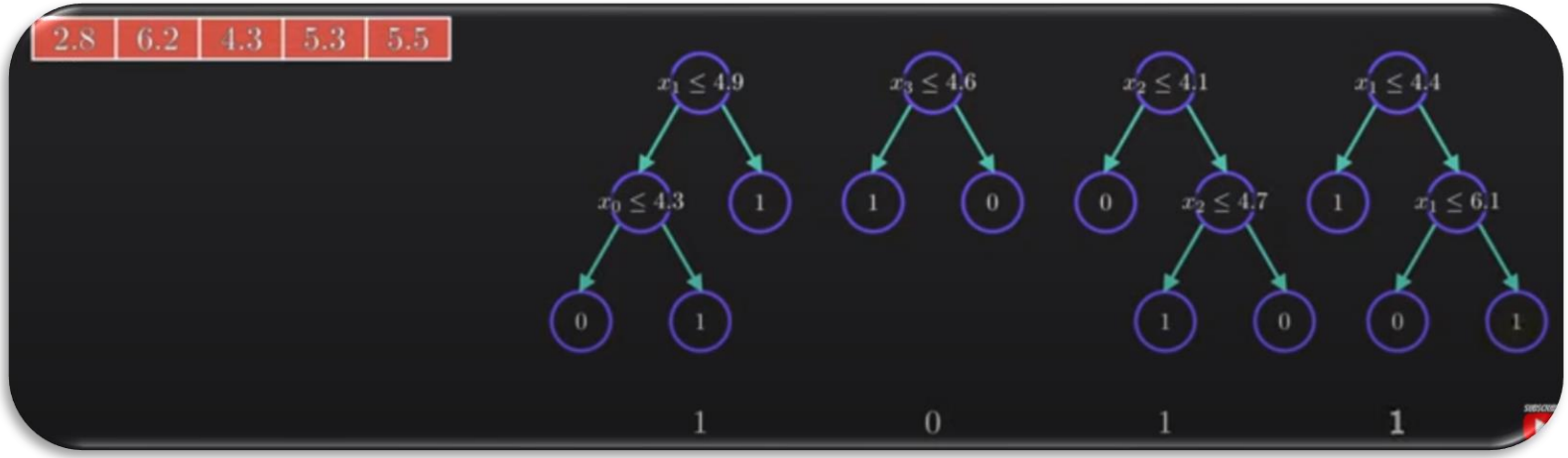
Faculdade
XPe

# Árvore # 4



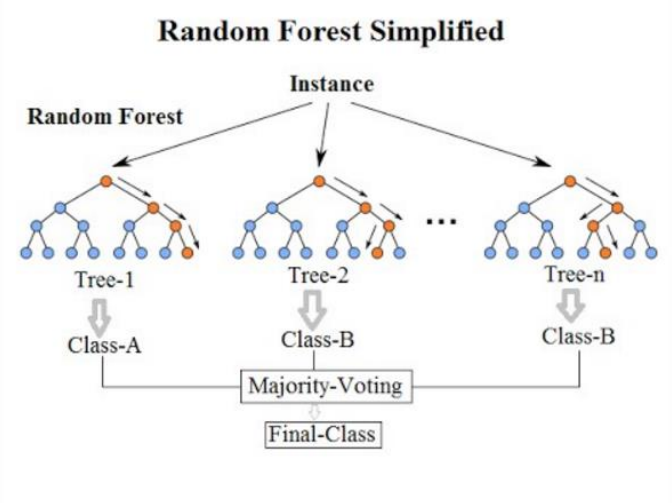
Faculdade XPe

# Como fazer uma previsão?



Faculdade XPe

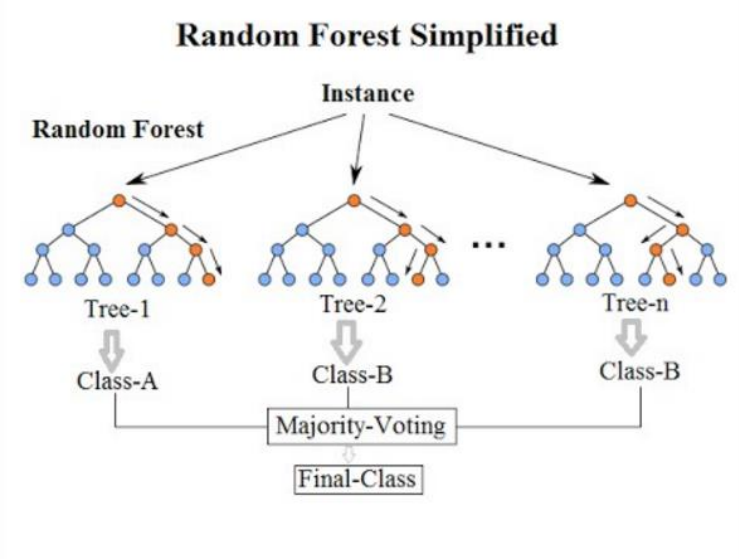
# Como fazer uma previsão?



Faculdade XPe

# Por que a floresta tende a diminuir a variância?

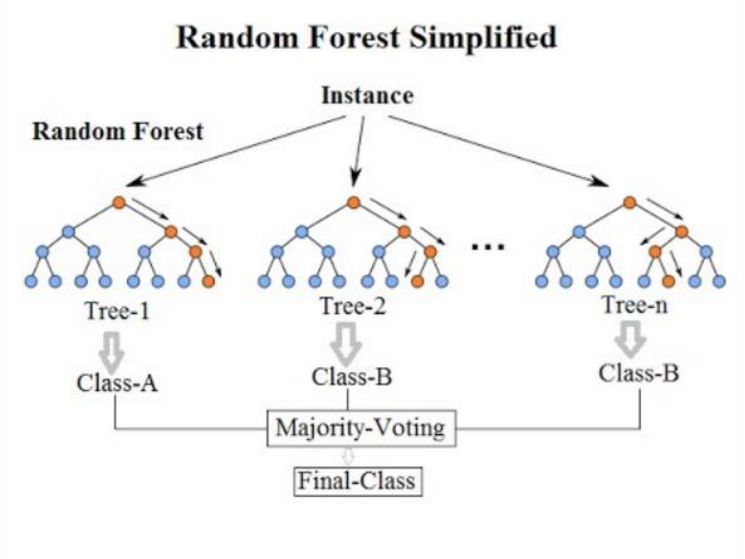
- Várias árvores diminuem a sensibilidade a registros e colunas específicas
- Menos *overfitting*!



Faculdade XPe

# Por que a floresta tende a diminuir a variância?

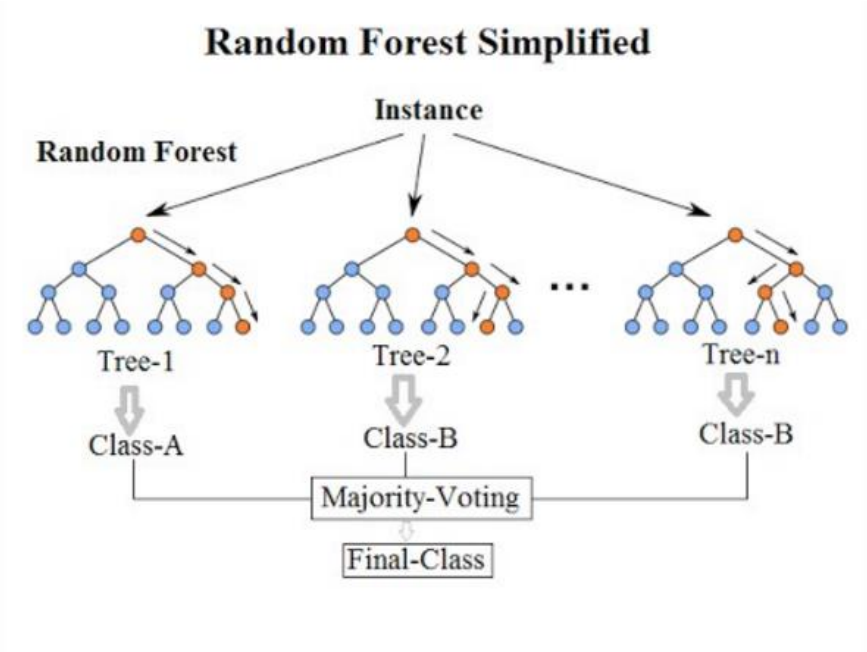
Fraquezas individuais
de cada árvore
tendem a se anular



Faculdade XPe

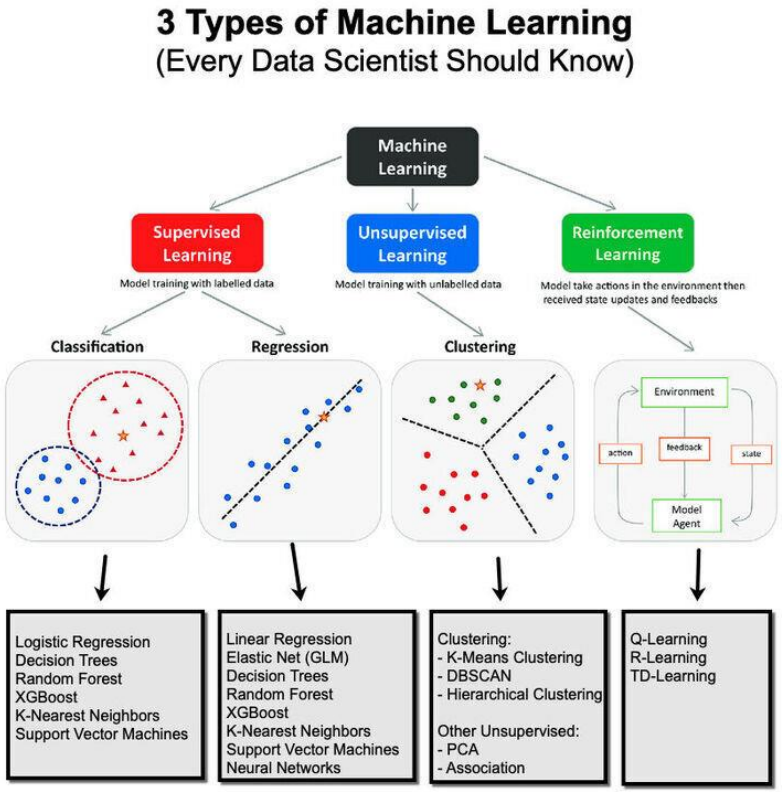
# Conclusão

Bem-vindo às *random* forests :-)



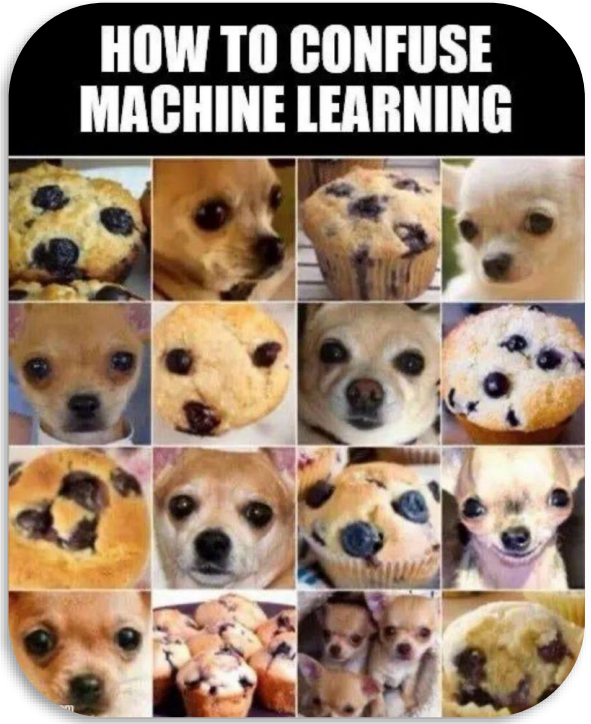
Faculdade XPe

# Os tipos de aprendizado de máquina



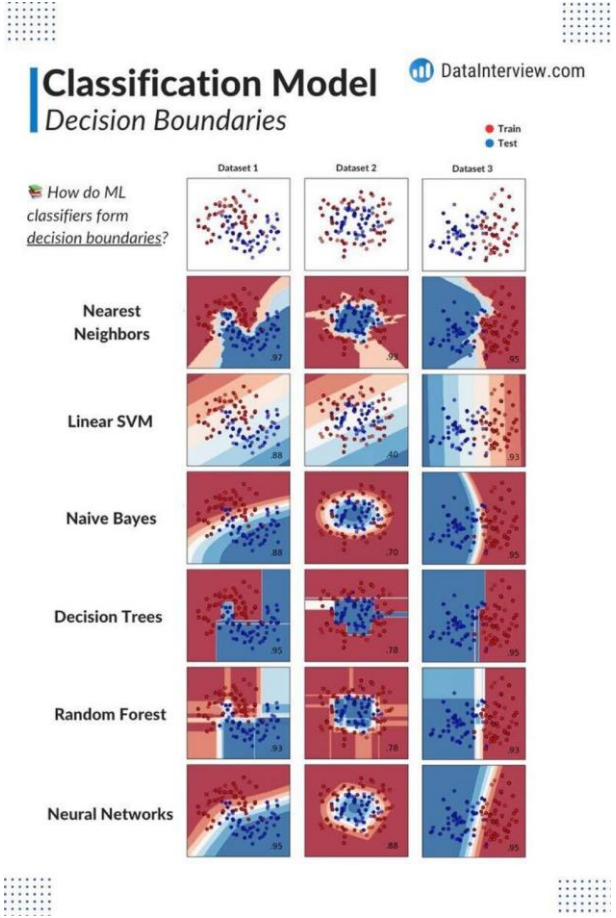
Faculdade XPe

# Como a máquina se confunde?



Faculdade XPe

# Modelos e fronteiras de decisão




Fundamentos de Machine Learning