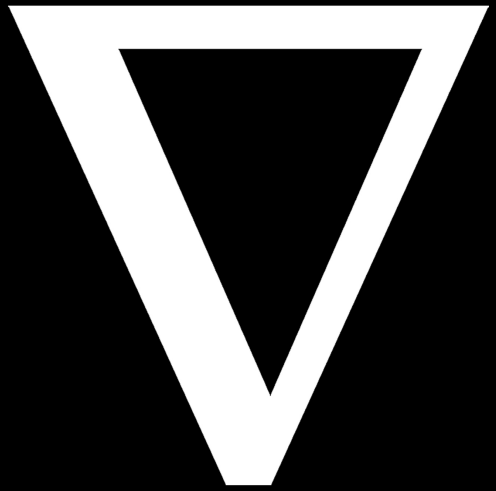
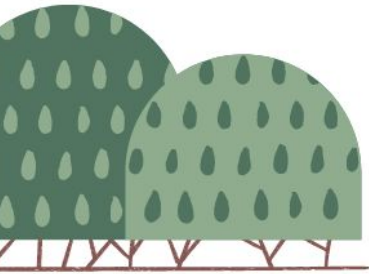


Árvores de Decisão e Random Forest

Árvores de Decisão

Árvores de Decisão

- Aprendizado Supervisionado
- Representação mais interpretável do conhecimento
- Hierarquia de decisões
- Utiliza a estratégia dividir para conquistar



Árvores de Decisão

Uma pessoa quer aprender sobre Árvores de Decisão

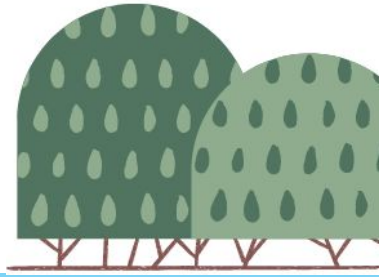
Presta atenção na aula do Data :)

Não presta atenção na aula :(

Faz uma declaração e toma uma decisão baseada se a declaração é verdadeira ou falsa

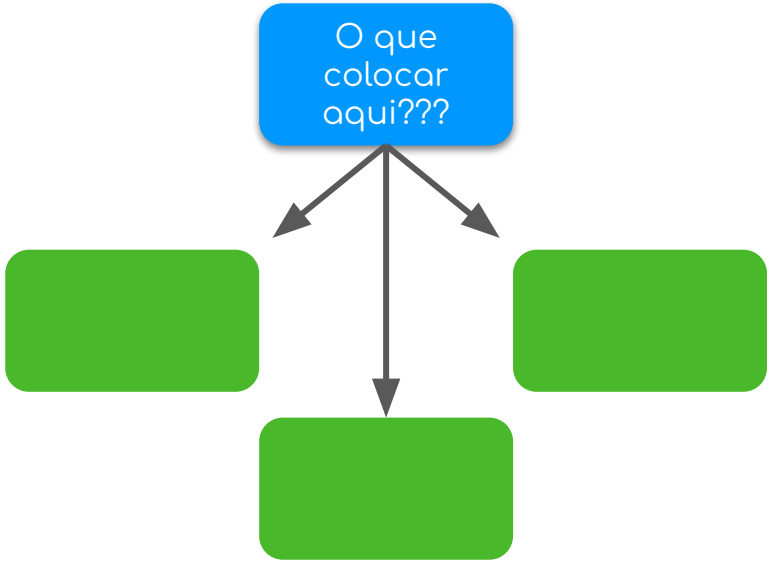
Como montar?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



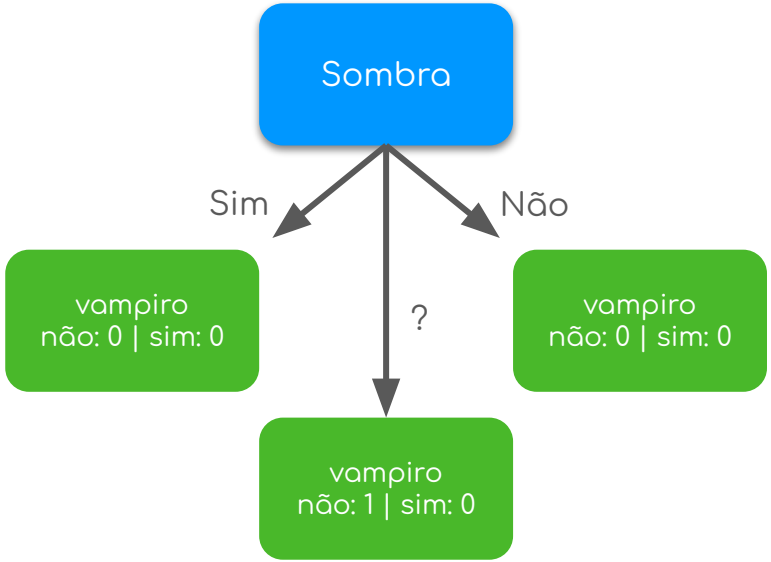
Como montar?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



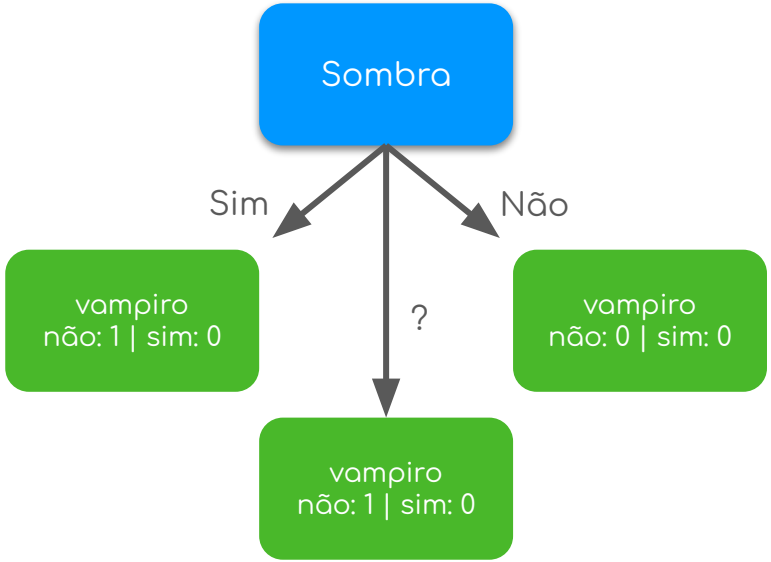
Quão bem sombra prevê?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



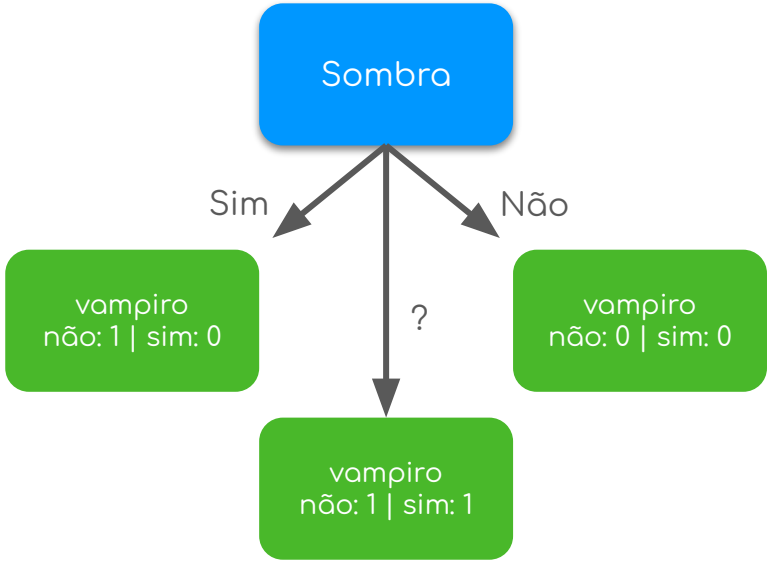
Quão bem sombra prevê?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



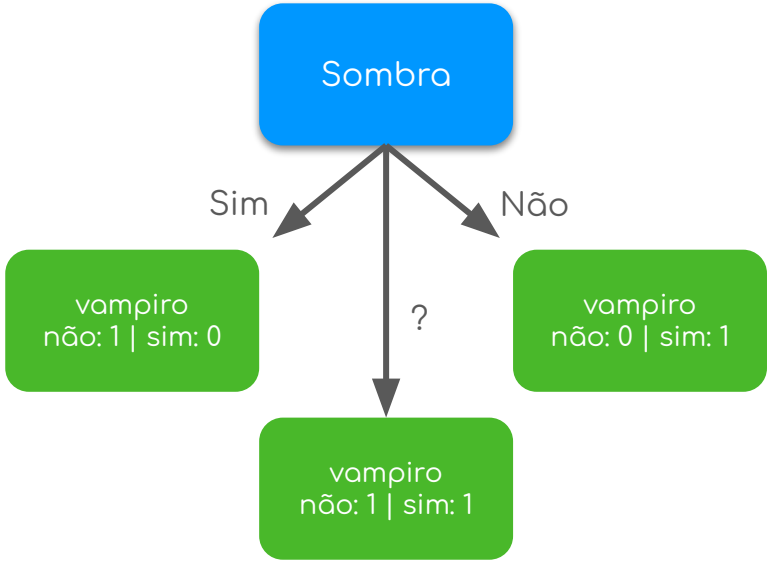
Quão bem sombra prevê?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



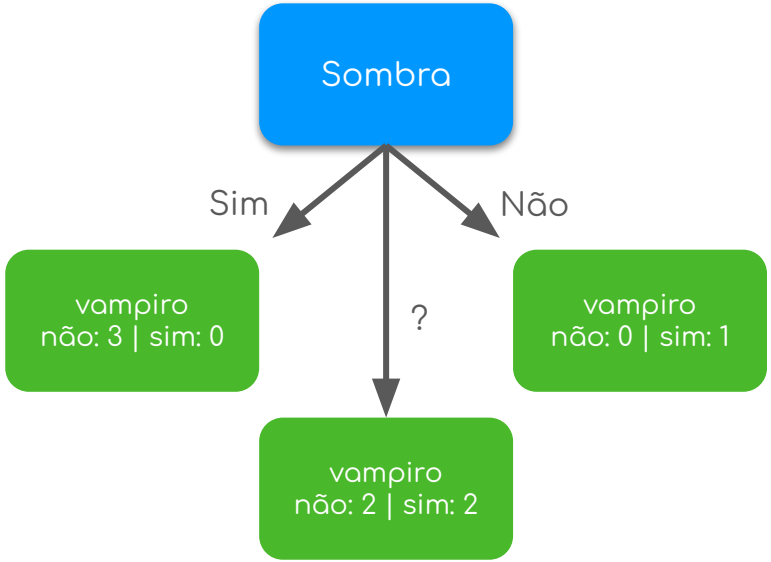
Quão bem sombra prevê?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



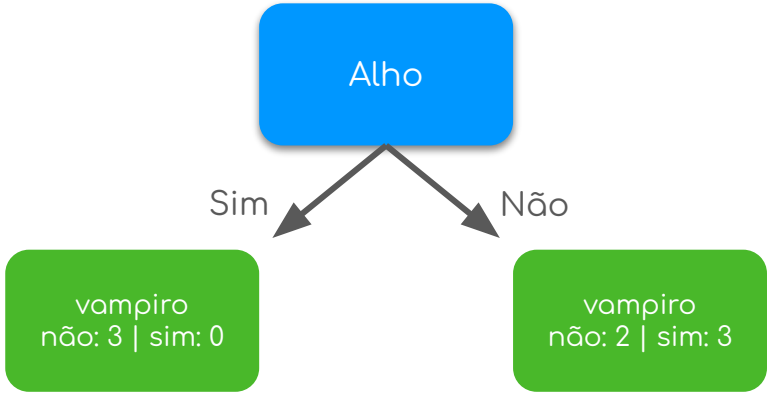
Quão bem sombra prevê?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



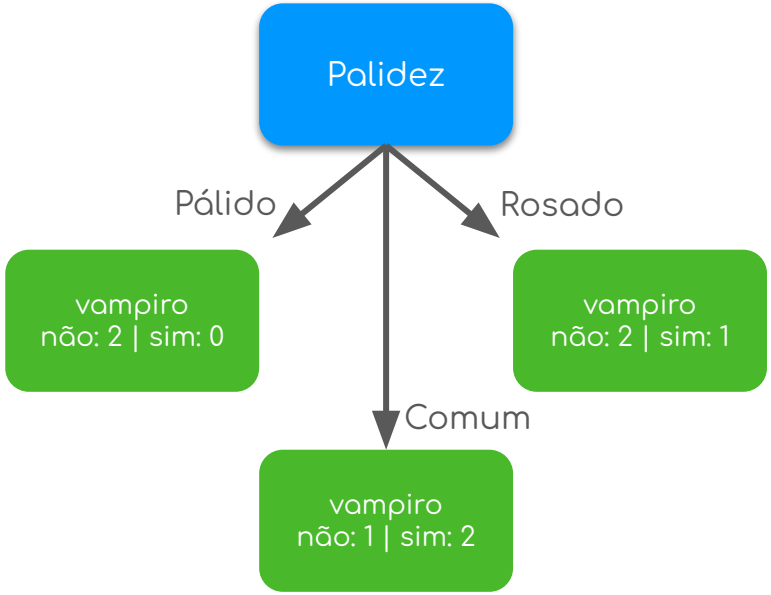
Quão bem alho prevê?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não

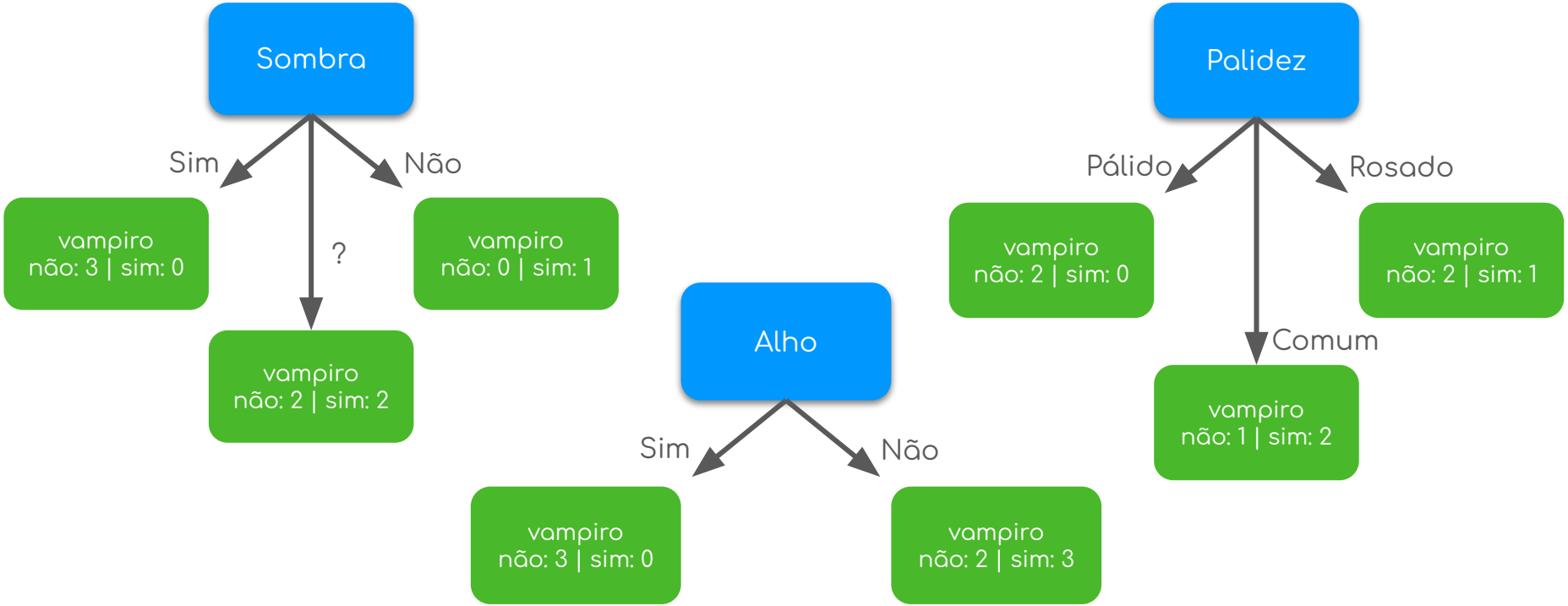


Quão bem palidez prevê?

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não

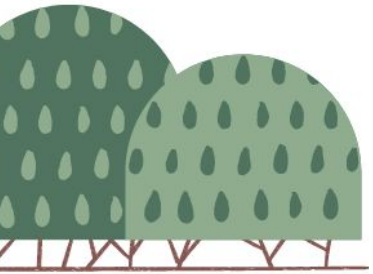


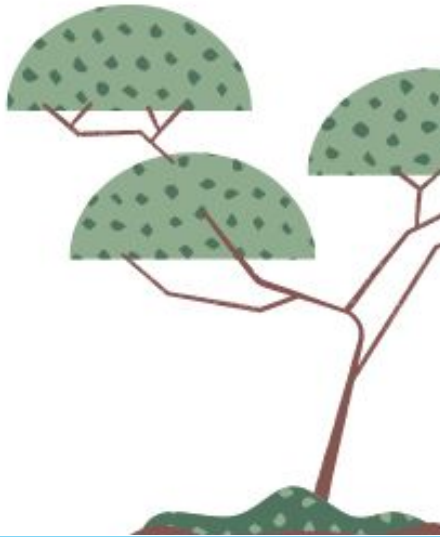
Comparativo



Comparativo

- Nenhum dos três faz um trabalho **perfeito** de predição!
- Como prosseguir?
- Como compará-los?



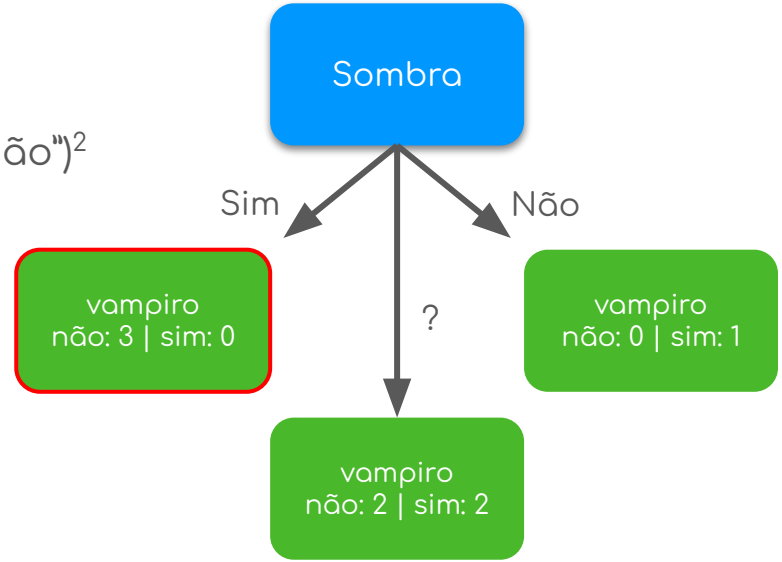


Quantificando impureza

- Entropia
- Ganho de Informação
- **Impureza Gini**
- Métodos numericamente similares

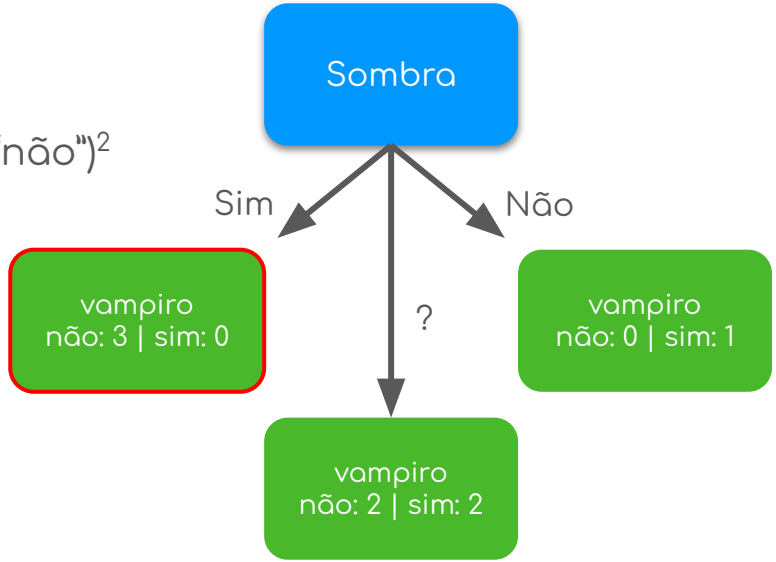
Quantificando impureza

Impureza Gini de um nó:
1 - (probabilidade de “sim”)^2 - (probabilidade de “não”)^2



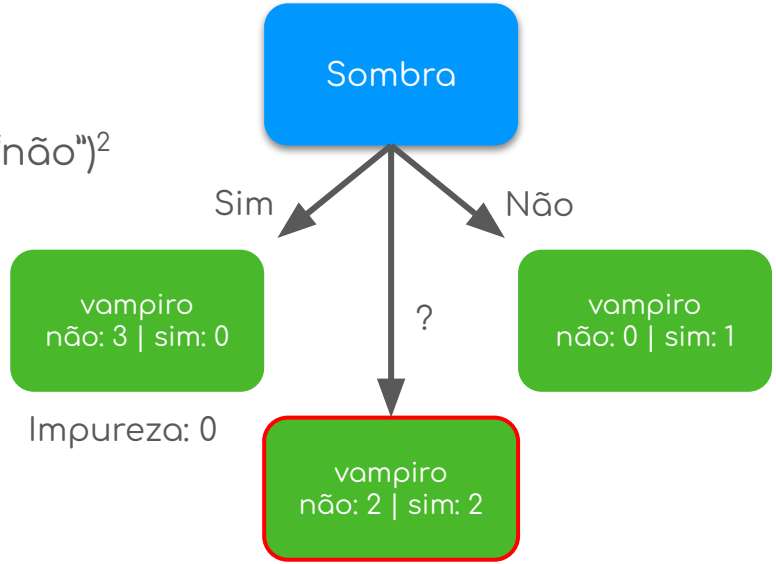
Quantificando impureza

Impureza Gini de um nó:
= 1 - (probabilidade de “sim”)^2 - (probabilidade de “não”)^2
= 1 - $(0/3)^2$ - $(3/3)^2$
= 1 - 0 - 1
= 0



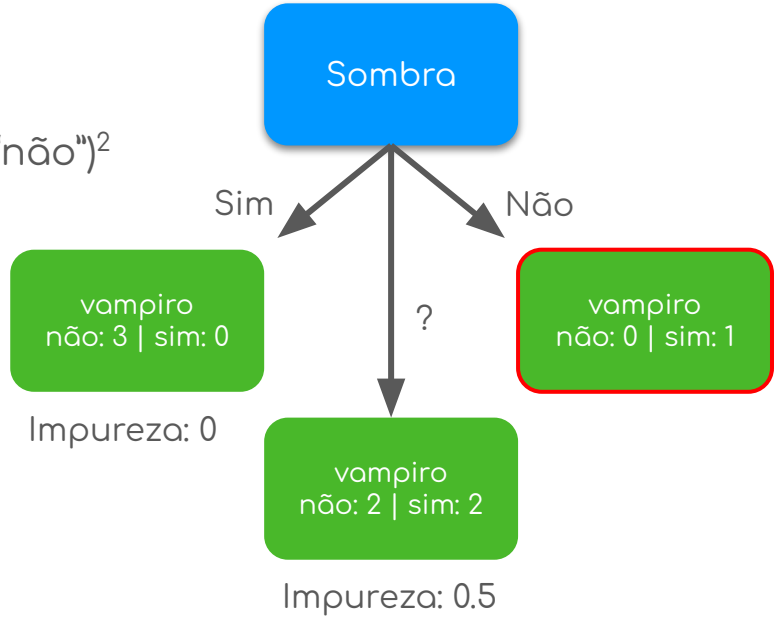
Quantificando impureza

Impureza Gini de um nó:
= 1 - (probabilidade de "sim")^2 - (probabilidade de "não")^2
= 1 - $(2/4)^2$ - $(2/4)^2$
= 1 - (¼) - (¼)
= (½)
= 0.5



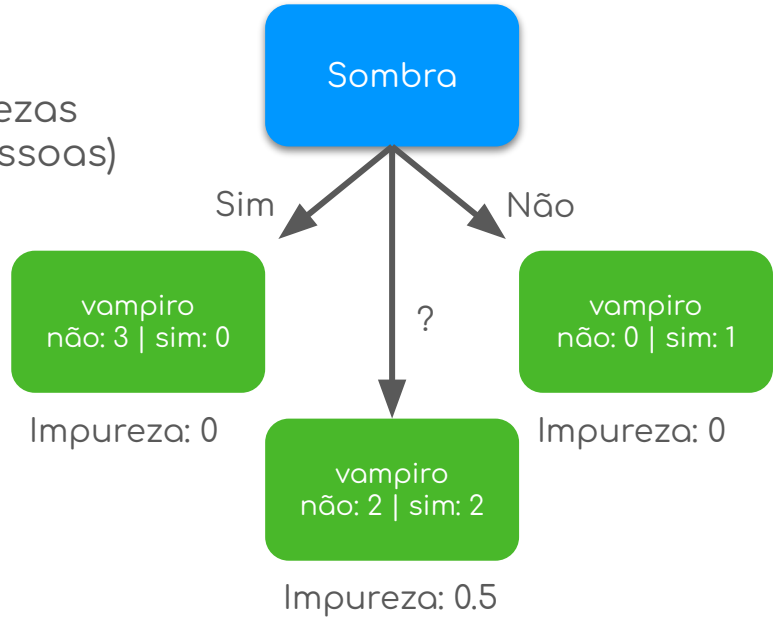
Quantificando impureza

Impureza Gini de um nó:
= 1 - (probabilidade de "sim")^2 - (probabilidade de "não")^2
= 1 - $(1/1)^2$ - $(0/1)^2$
= 1 - 1 - 0
= 0

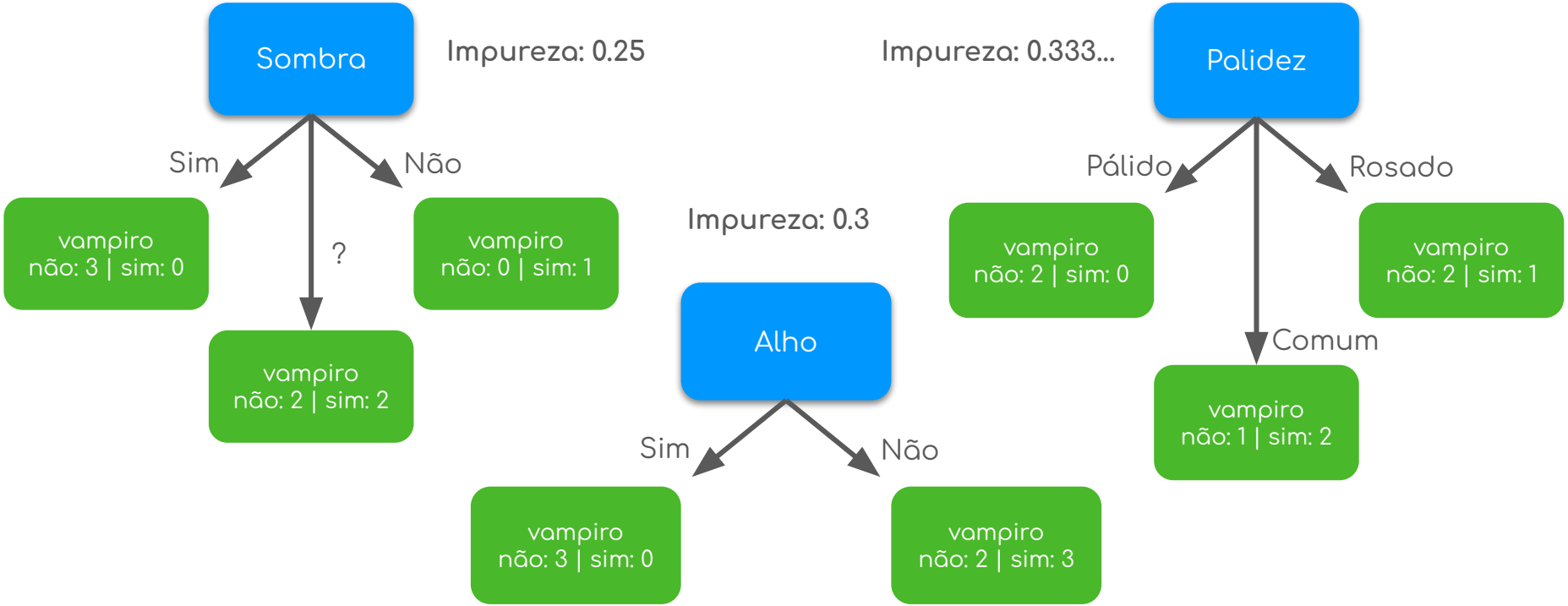


Quantificando impureza

Impureza Gini total: média ponderada das impurezas
Pesos: (quantidade de pessoas no nó/total de pessoas)
= (⅜) * 0 + (4/8) * 0.5 + (⅛) * 0
= 0 + (¼) + 0
= (¼)
= 0.25

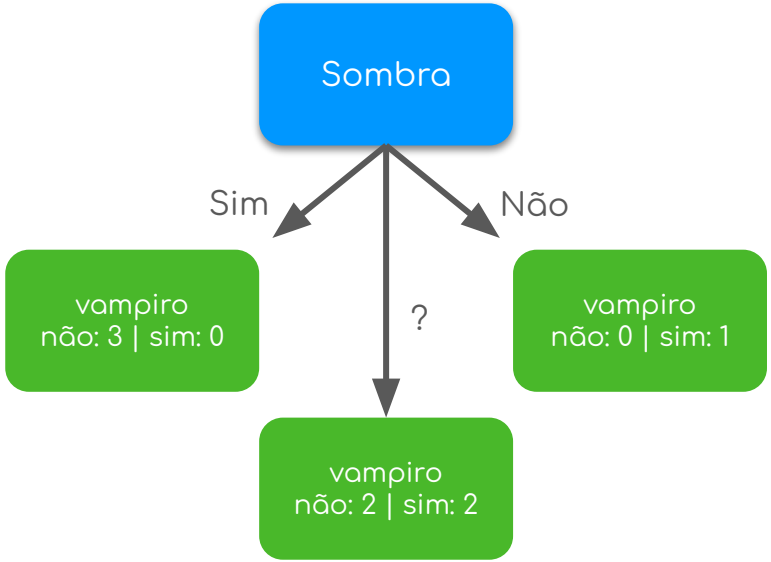


Comparativo



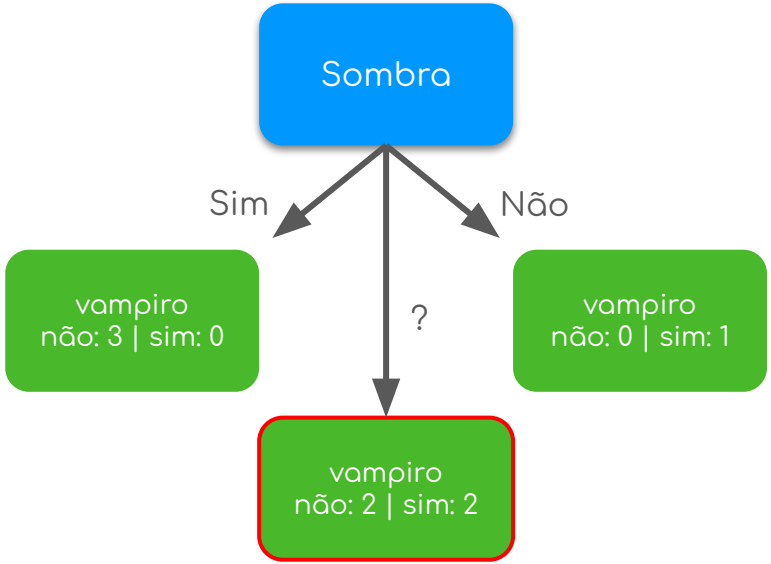
Sombra prevê melhor

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



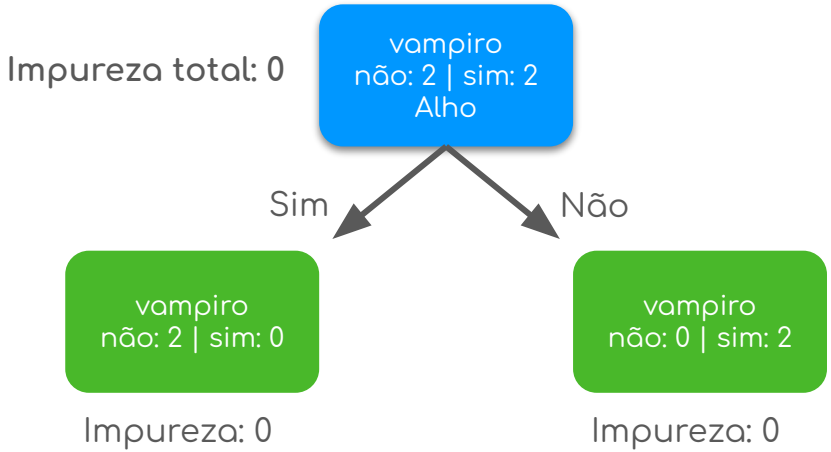
Sombra prevê melhor

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



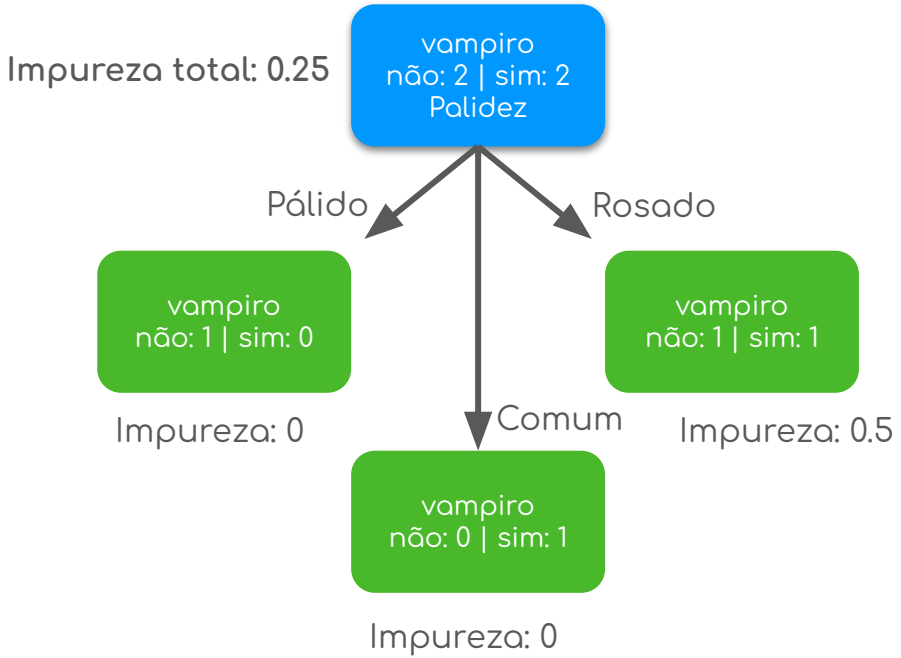
Fazemos tudo novamente :)

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não

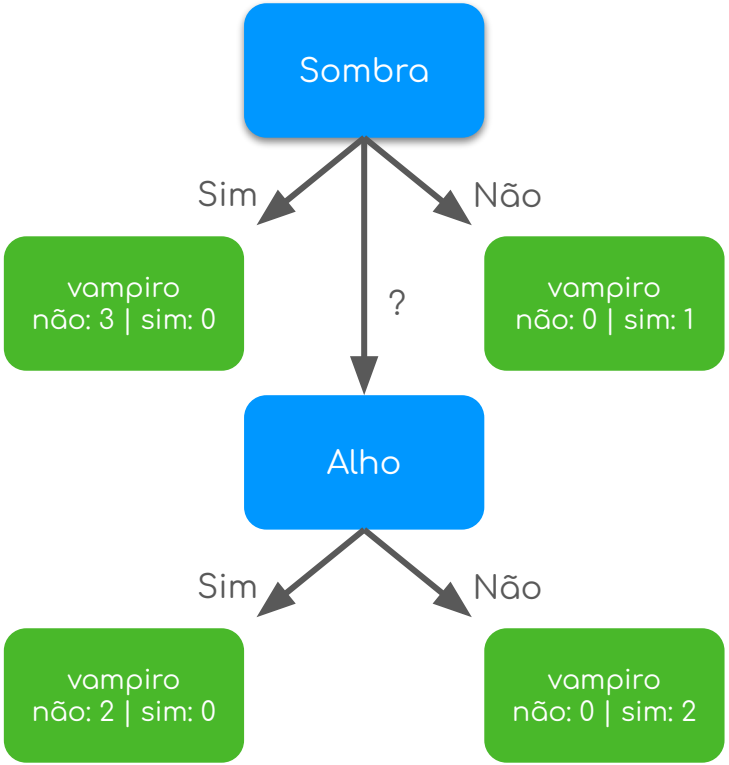


Fazemos tudo novamente :)

Sombra	Alho	Palidez	Vampiro
?	Sim	Pálido	Não
Sim	Sim	Rosado	Não
?	Não	Rosado	Sim
Não	Não	Comum	Sim
?	Não	Comum	Sim
Sim	Não	Pálido	Não
Sim	Não	Comum	Não
?	Sim	Rosado	Não



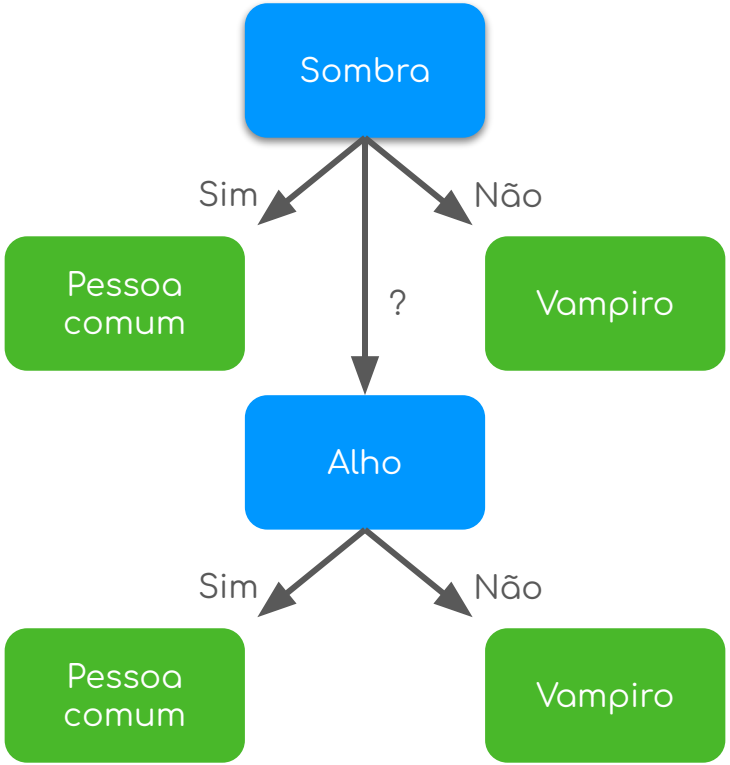
Árvore quase pronta



É preciso atribuir valores de saída para cada nó folha

De maneira geral a saída é a categoria com a maior quantidade de “votos”

Árvore de Decisão



Quantificando impureza

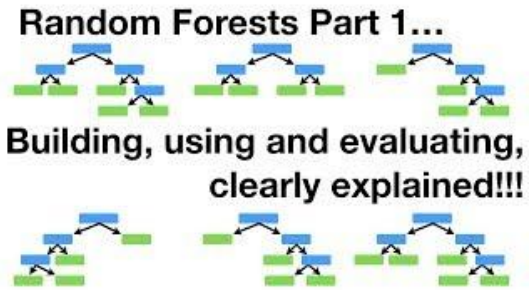
- Com a árvore pronta se alguém novo aparece, consigamos prever se é um vampiro ou não!
- Mas, ...

Métodos de Conjunto de Árvores

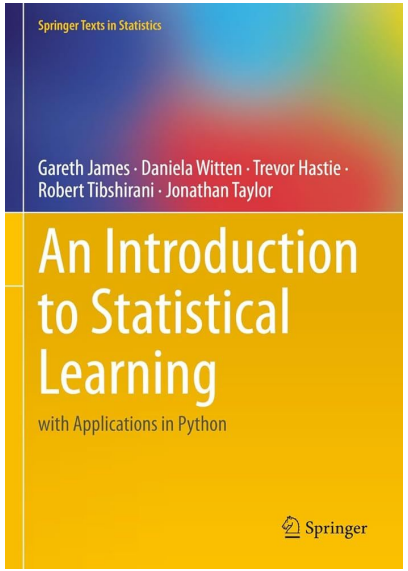
Por que precisamos combinar árvores?

- **Decision Trees** são fáceis de construir, fáceis de usar e fáceis de interpretar, PORÉM são **imprecisas**.
- Além disso, sofrem com uma **alta variância**.
- Então, combinamos muitos **modelos simples** para obter um **modelo único** e potencialmente poderoso.

Referências



https://www.youtube.com/watch?v=J4Wdy0Wc_xQ&ab_channel=StatQuestwithJoshStarmer



https://www.statlearning.com/

- https://www.youtube.com/watch?v=xWhPwHZF4c0&ab_channel=StanfordOnline
- https://www.youtube.com/watch?v=tjy0yL1rRRU&ab_channel=DataMListic
- https://www.youtube.com/watch?v=GM3CDQfQ4sw&ab_channel=Udacity
- https://towardsdatascience.com/introduction-to-boosted-trees-2692b6653b53
- https://aws.amazon.com/pt/what-is/boosting/
- https://towardsmachinelearning.org/boosting-algorithms/
- https://mathchi.medium.com/weak-learners-strong-learners-for-machine-learning-e73e32f86ebd

Métodos

- Bagging
- Random Forest
- Boosting
- Bayesian Additive Regression Trees

Métodos

- Bagging
- Random Forest
- Boosting
- Bayesian Additive Regression Trees

* O maior foco da aula será em Bagging/Random Forest

Métodos

- Bagging
- Random Forest
- Boosting
- Bayesian Additive Regression Trees

* O maior foco da aula será em Bagging/Random Forest

* Alguns desses métodos não são específicos para árvores, mas estão presentes nesse contexto.

Bagging e Random Forest

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Bagging e Random Forest

1. Criar um Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease

Bagging e Random Forest

1. Criar um Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes

Bagging e Random Forest

1. Criar um Bootstrapped Dataset

Chest Pain	Good Blood Circ	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No

Bagging e Random Forest

1. Criar um Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes

Bagging e Random Forest

1. Criar um Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Bagging e Random Forest

2. Criar uma **decision tree** com um número **M** de features(ou colunas).

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

* Este é o ponto que diferencia o bagging de uma random forest.

Bagging e Random Forest

2. Criar uma **decision tree** com um número **M** de features(ou colunas).

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Sendo **ρ** o número total de features no dataset.

Bagging:

M = ρ

Bagging e Random Forest

2. Criar uma **decision tree** com um número **M** de features(ou colunas).

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Sendo **ρ** o número total de features no dataset.

Random Forest:

$M < \rho$

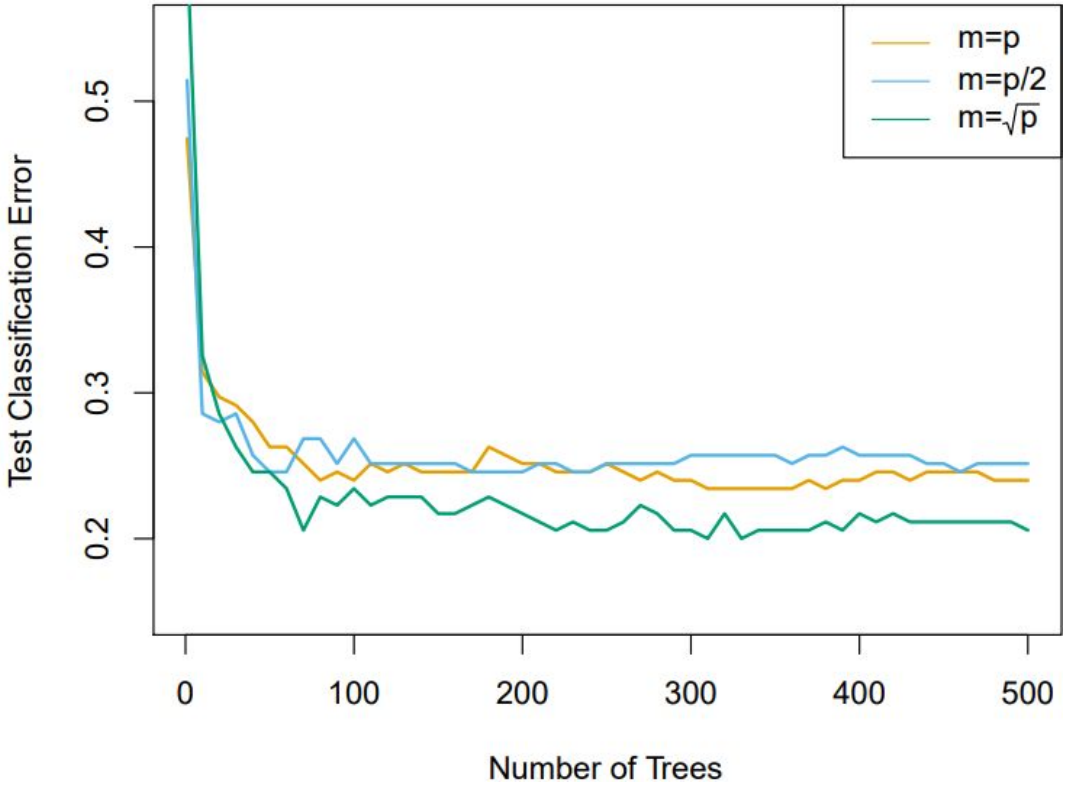
Bagging e Random Forest

2. Criar uma **decision tree** com um número **M** de features(ou colunas).

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Obs: A escolha para M é arbitrária, mas geralmente é utilizado $M = \sqrt{p}$

Bagging e Random Forest



Bagging e Random Forest

Por que usar um subconjunto das features?

"Suponha que haja um preditor **muito forte** no conjunto de dados, juntamente com vários outros preditores moderadamente fortes. Então, na coleção de "bagged trees", a maioria ou todas as árvores **usarão esse forte preditor na divisão superior**. Consequentemente, todas as árvores ensacadas **serão bastante semelhantes entre si.**"

" isso significa que o bagging **não levará a uma redução substancial na variância** sobre uma única árvore neste cenário."

Bagging e Random Forest

2. Criar uma **decision tree** com um número **M** de features(ou colunas).

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Nesse caso, foi escolhido aleatoriamente **Good Blood Circulation** e **Blocked Arteries** como candidatos para a raiz da árvore.

Bagging e Random Forest

2. Criar uma **decision tree** com um número **M** de features(ou colunas).

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

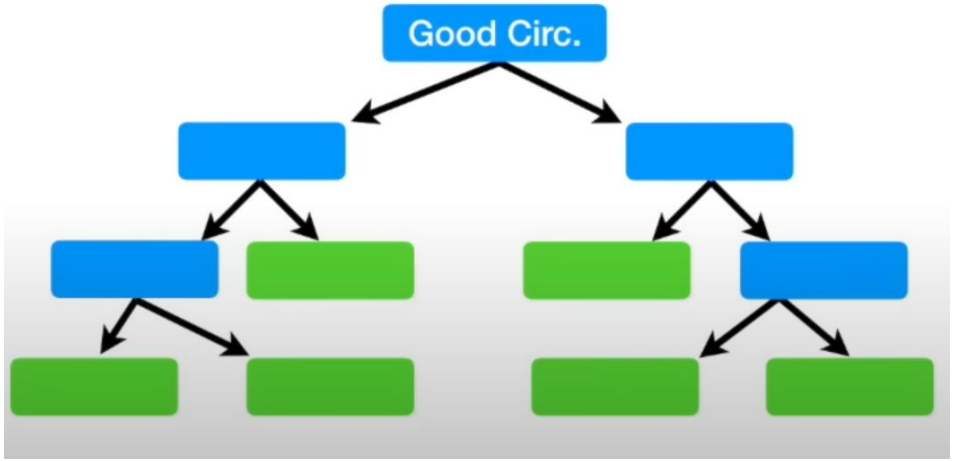
Bagging e Random Forest

2. Criar uma **decision tree** com um número **M** de features(ou colunas).

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Bagging e Random Forest

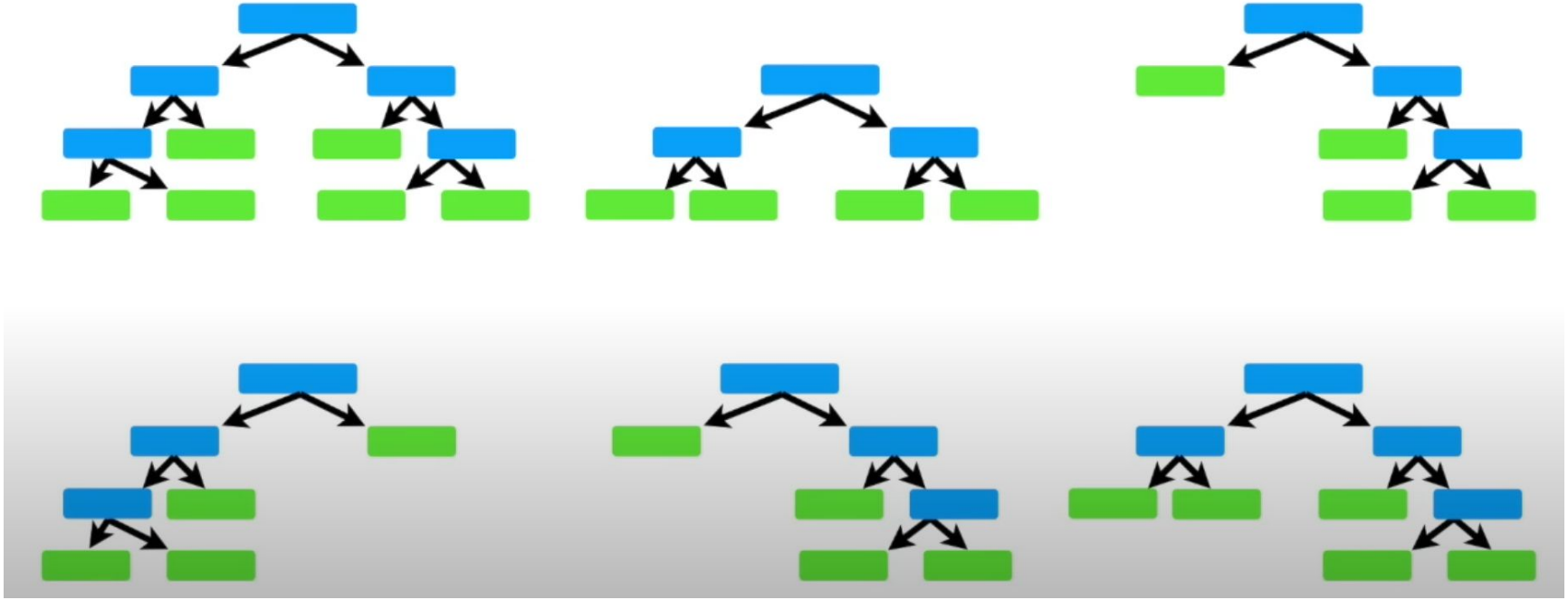
2. Criar uma **decision tree** com um número **M** de features(ou colunas).



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Bagging e Random Forest

3. Repete o processo várias vezes



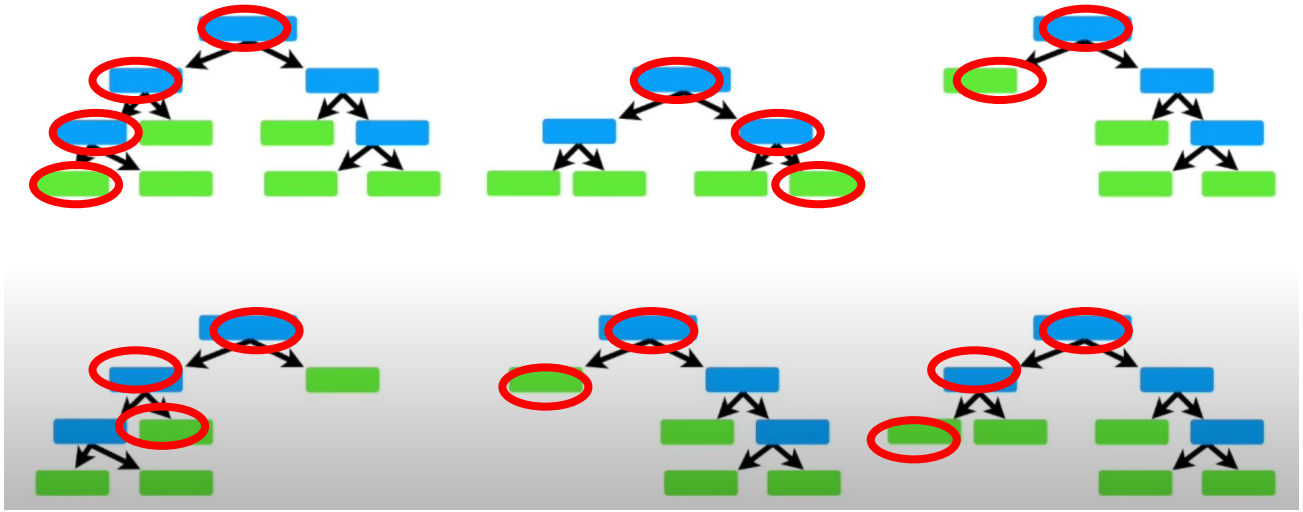
Bagging e Random Forest

Como usar?

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

Bagging e Random Forest

Como usar?



Heart Disease	
Yes	No
5	1

Bagging e Random Forest

Como Avaliar?

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Quando estávamos construindo o Bootstrapped Dataset, essa foi uma linha que ficou de fora.

Bagging e Random Forest

Como Avaliar?

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Out-of-Bag Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	No	210	No

Bagging e Random Forest

Como Avaliar?

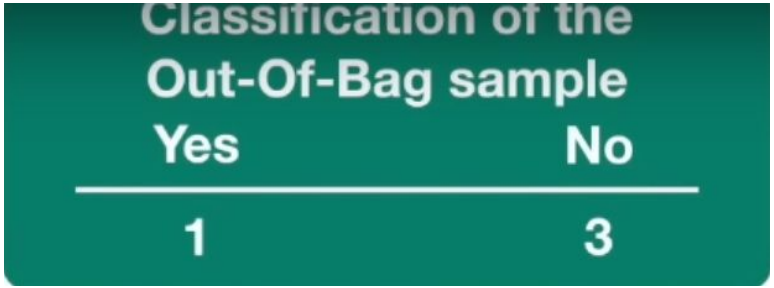
Out-of-Bag Dataset

- Neste caso, possui apenas uma entrada;

- Cerca de 1/3 das entradas não são selecionadas e vão para o out-of-bag dataset.

Bagging e Random Forest

Como Avaliar?



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes

Bagging e Random Forest

Como Avaliar?

Classification of the Out-Of-Bag sample	
Yes	No
4	0

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes

Bagging e Random Forest

Como Avaliar?

Classification of the Out-Of-Bag sample	
Yes	No
3	1

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes

Bagging e Random Forest

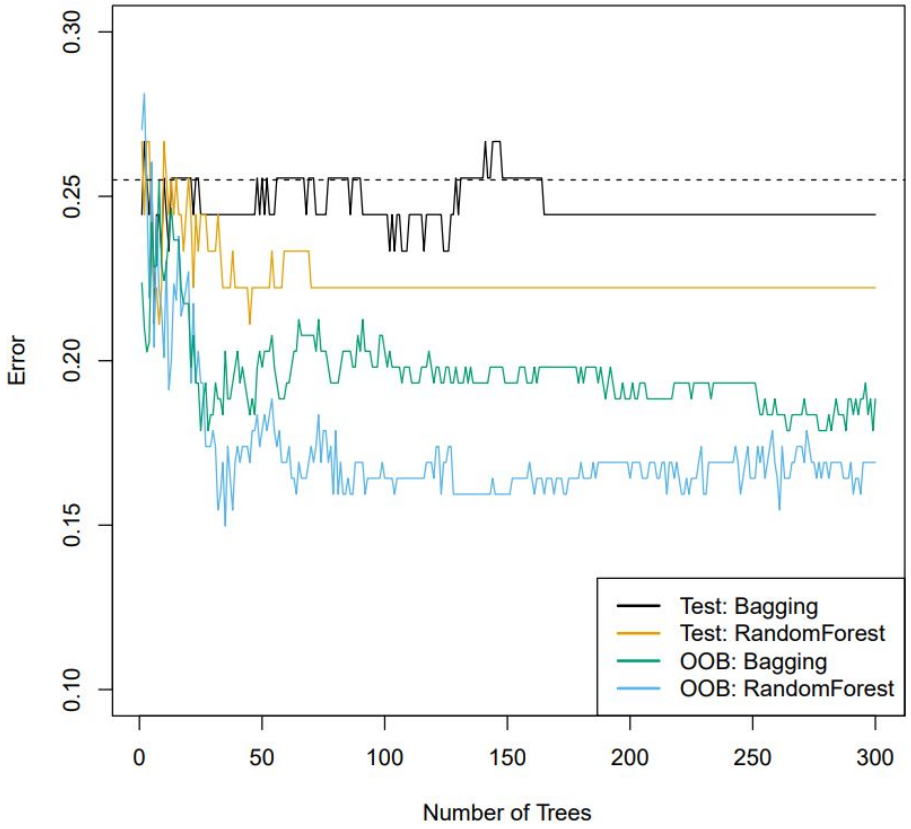
Como Avaliar?

Classification of the Out-Of-Bag sample	
Yes	No
3	1

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes

A proporção de amostras out-of-bag que estão incorretas é chamada de **Out-of-Bag Error**

Bagging e Random Forest



- O erro de uma única árvore é muito grande;
- À medida que aumenta o número de árvores, o erro diminui e depois estabiliza;
- Usar um valor muito grande de B não levará ao overfitting.

Boosting

“Eles (Random Forest) não conseguem lidar com erros (se houver) criados por suas árvores de decisão individuais. Devido ao aprendizado paralelo, **se uma árvore de decisão cometer um erro, todo o modelo de floresta aleatória comete esse erro.** “

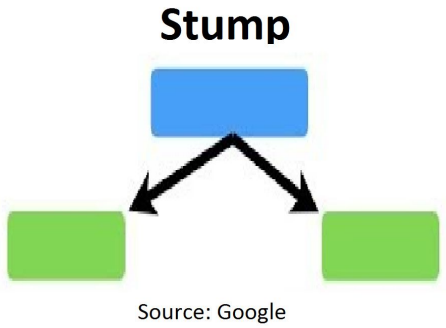
https://towardsdatascience.com/introduction-to-boosted-trees-2692b6653b53

Boosting

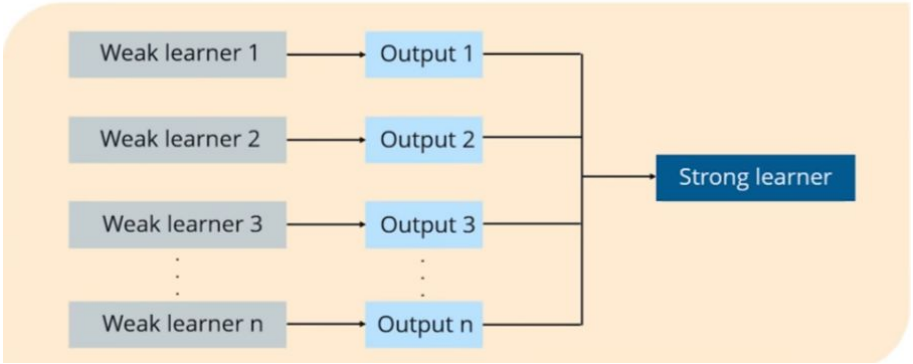
- Novas árvores são formadas considerando os erros das árvores nas rodadas anteriores.
- Aprendizado Sequencial
- Processo iterativo
- São usados principalmente para reduzir **viés** e **variância**.

Boosting

- Transforma árvores de decisão fracas (chamadas de **weak learners**) em **strong learners**;

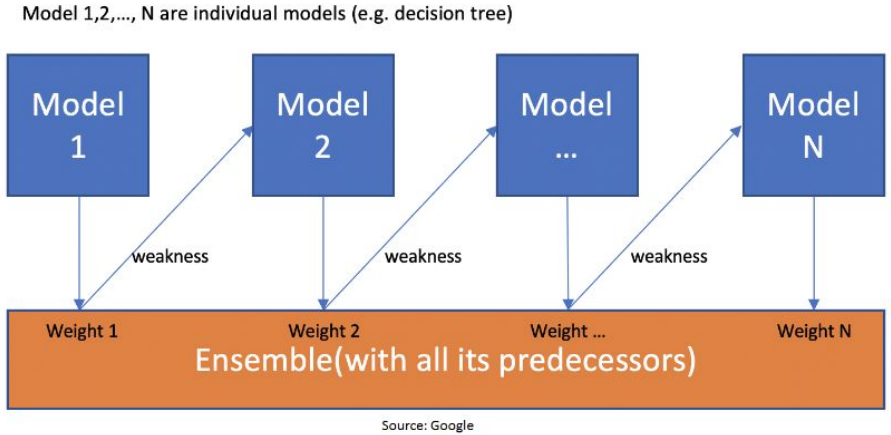


Source: Google

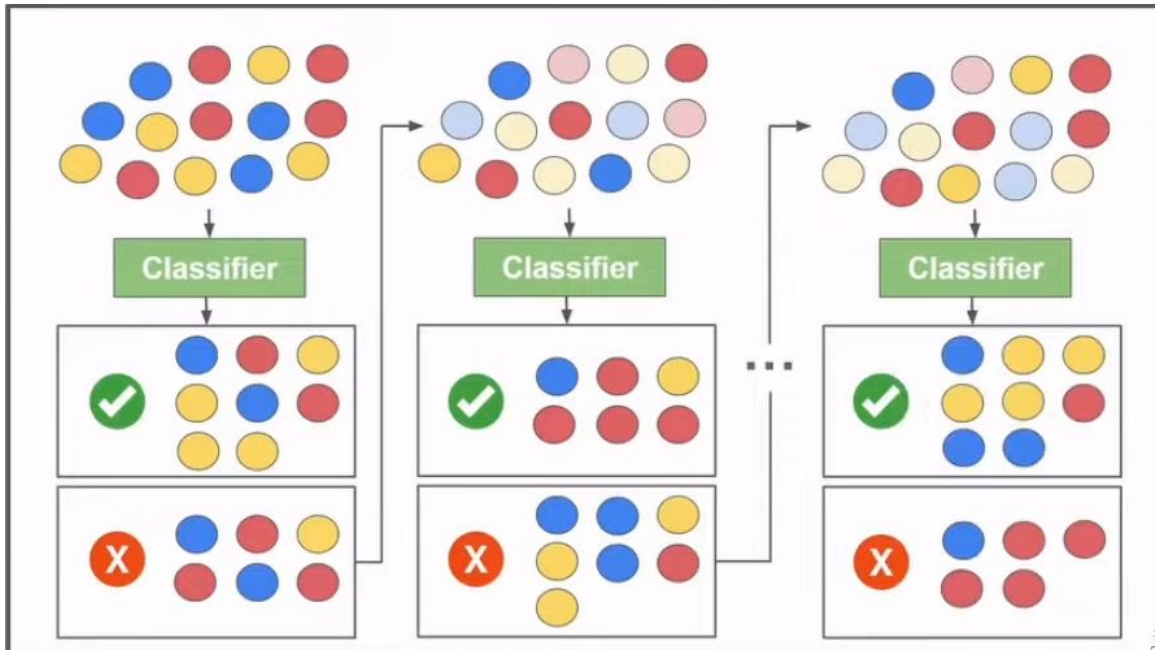


Boosting

- Transforma árvores de decisão fracas (chamadas de **weak learners**) em **strong learners**;



Boosting



https://www.youtube.com/watch?v=tjy0yL1rRRU&ab_channel=DataMListic

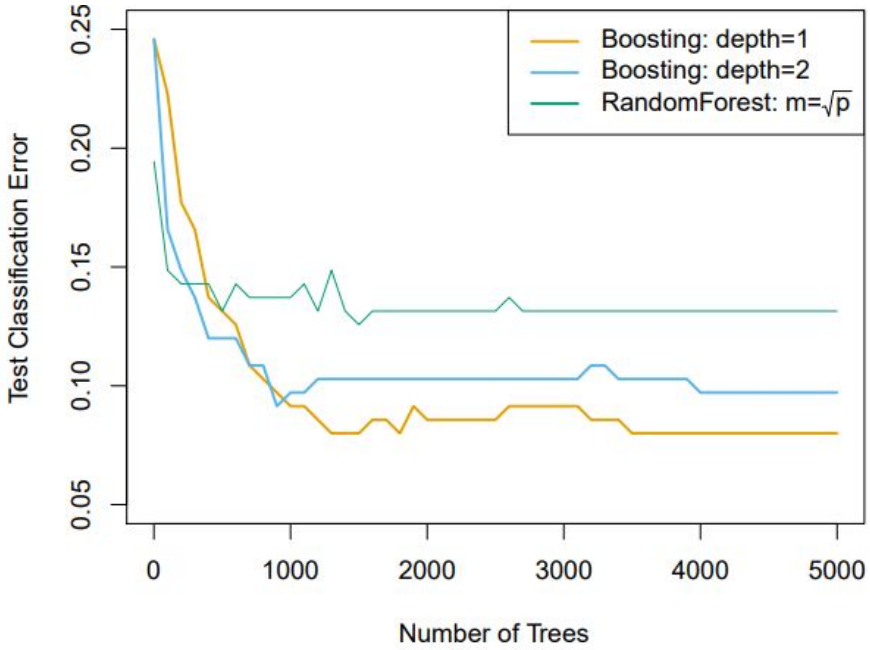
Boosting

Algoritmos famosos:

- AdaBoost (Adaptive Boosting)
- Gradient Boosting
- XGBoost (Extreme Gradient Boosting)
- LightGBM (Light Gradient Boosting Machine)
- CatBoost (Categorical Boosting)

Boosting

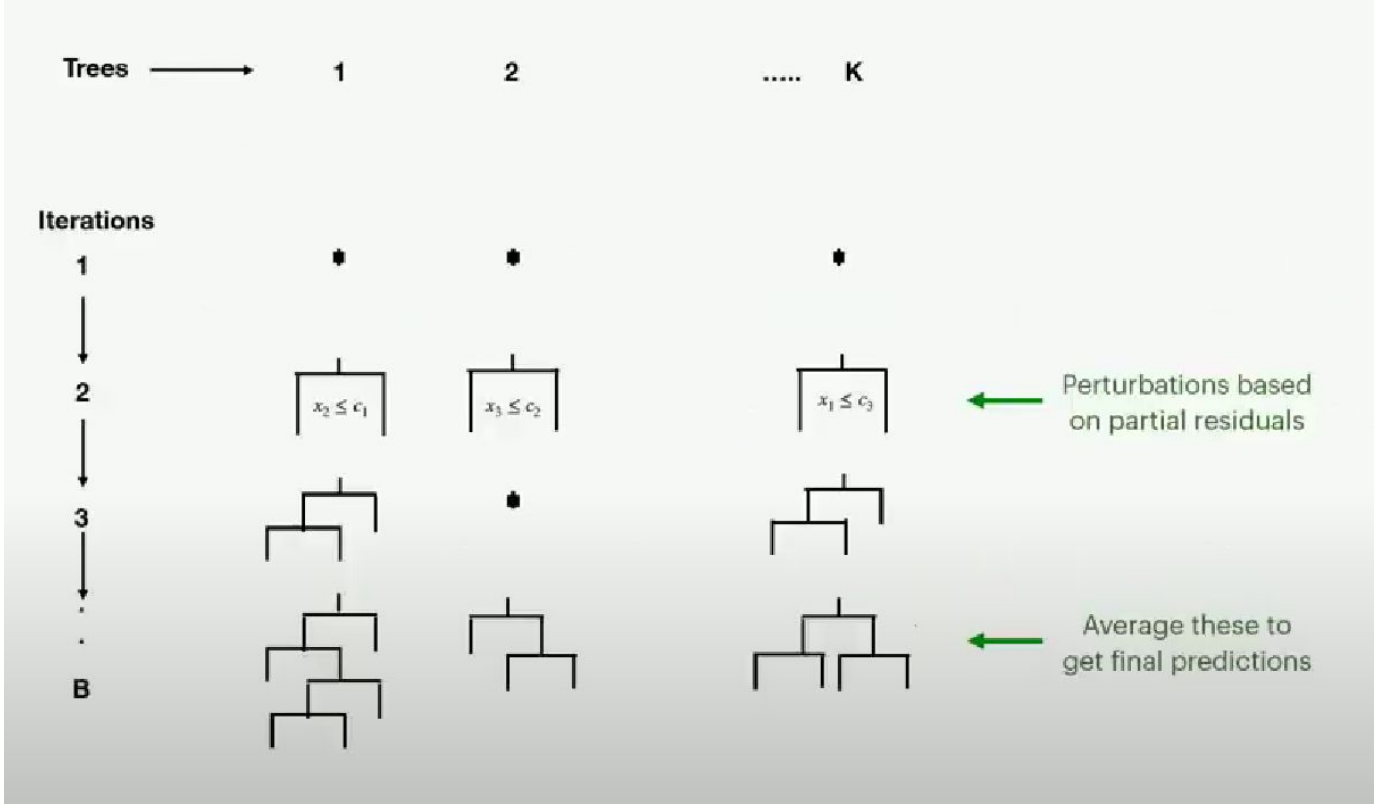
- Pode sofrer **overfitting**, então cuidado com seus hiperparâmetros;



Bayesian Additive Regression Trees (BART)

“O BART está relacionado a ambas as abordagens: cada árvore é construída de **forma aleatória como em bagging e florestas aleatórias**, e cada árvore tenta **capturar sinais ainda não contabilizados pelo modelo atual, como no boosting**.“

Bayesian Additive Regression Trees (BART)



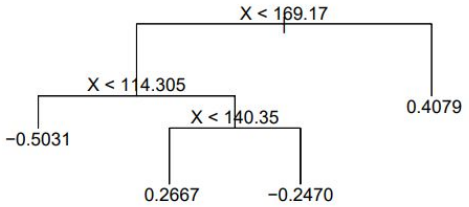
Bayesian Additive Regression Trees (BART)

Há dois componentes para esta perturbação:

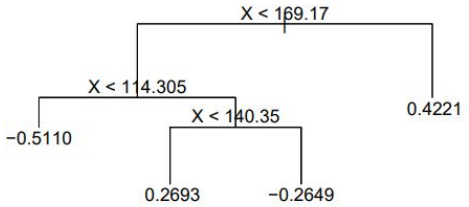
1. Podemos mudar a estrutura da árvore **adicionando ou podando** ramos(branch).
2. Podemos **alterar a previsão** em cada nó terminal da árvore

Bayesian Additive Regression Trees (BART)

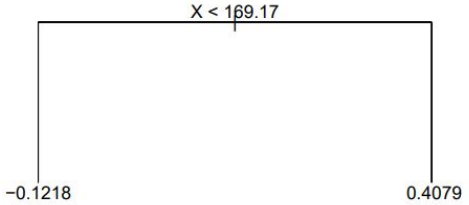
(a): $\hat{f}_k^{b-1}(X)$



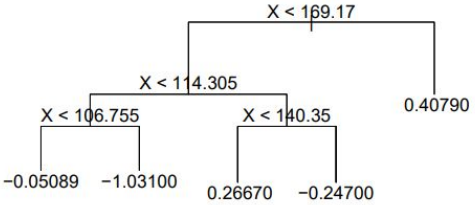
(b): Possibility #1 for $\hat{f}_k^b(X)$



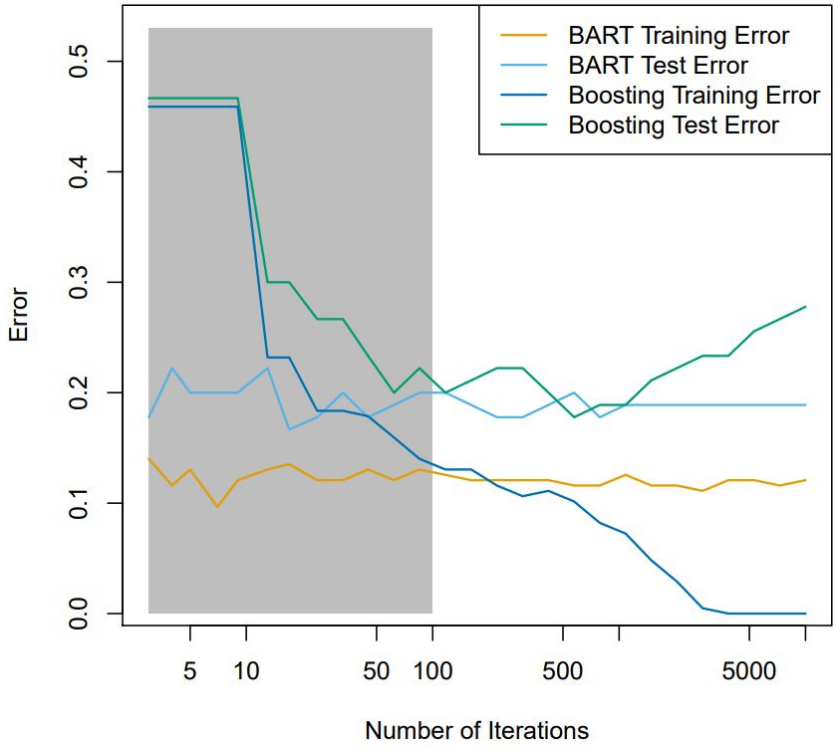
(c): Possibility #2 for $\hat{f}_k^b(X)$



(d): Possibility #3 for $\hat{f}_k^b(X)$



Bayesian Additive Regression Trees (BART)



Dúvidas?