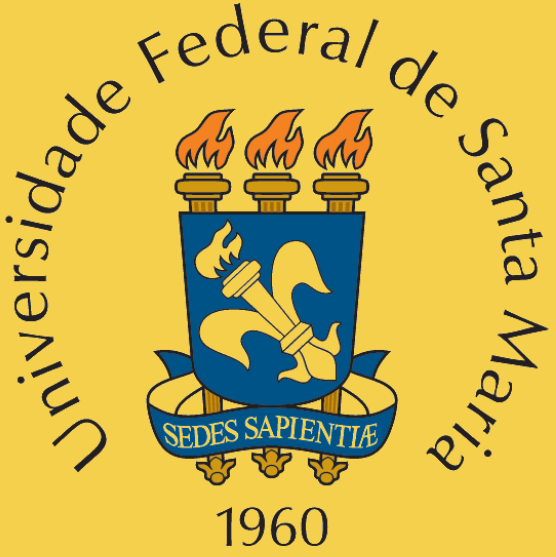

Universidade Federal de Santa Maria
SEDES SAPIENTIÆ
1960


UFVJM
1953
SAPIENTIA ET LIBERTAS

Reconhecimento de padrões e aprendizagem computacional

Métodos Ensemble

Métodos de Ensemble

Ensemble pode ser definido como Grupo

- Envolvem agrupar modelos preditivos de modo a melhorar a precisão e a estabilidade do modelo.

- São conhecidos por impulsionar e aprimorar os modelos baseados em árvore.

Métodos de Ensemble

Viés x Variância

- Como qualquer outro modelo, um modelo baseado em árvore também sofre com viés e variância.

Viés: ‘o quanto em média os valores previstos são diferentes dos valores reais’.

Variância: ‘o quão diferentes serão as previsões do modelo num mesmo ponto se diferentes amostras forem tomadas da mesma população’.

Métodos de Ensemble

Viés x Variância

- Suponha que você montou uma árvore pequena, obtendo um modelo com baixa variância e viés elevado.

Como então equilibrar o trade-off entre viés e variância?

- Normalmente, à medida que você aumenta a complexidade de seu modelo, você verá uma redução no erro de previsão devido ao viés mais baixo no modelo.

À medida que você continuar tornando o modelo mais complexo ele começará a sofrer com a variância.

Métodos de Ensemble

Viés x Variância

- Suponha que você montou uma árvore pequena, obtendo um modelo com baixa variância e viés elevado.

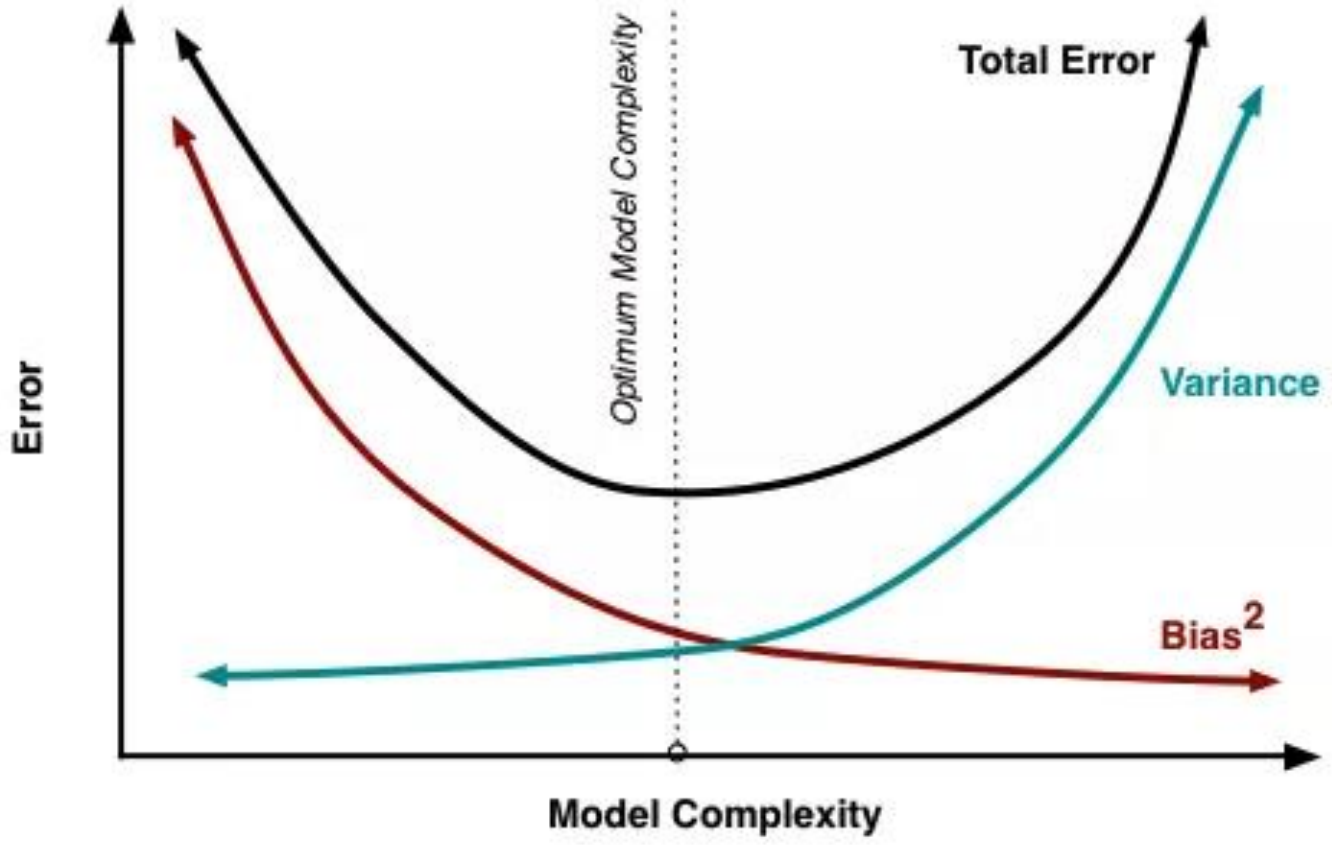
Como então equilibrar o trade-off entre viés e variância?

- Um modelo ótimo deve manter o equilíbrio entre estes dois tipos de erros. Isto é conhecido como a gestão de ‘trade-off’ entre erros de variância e viés.

Aprendizagem por ‘ensemble’ é uma maneira de analisar esse ‘trade-off’.

Métodos de Ensemble

Viés x Variância



Métodos de Ensemble

Bagging

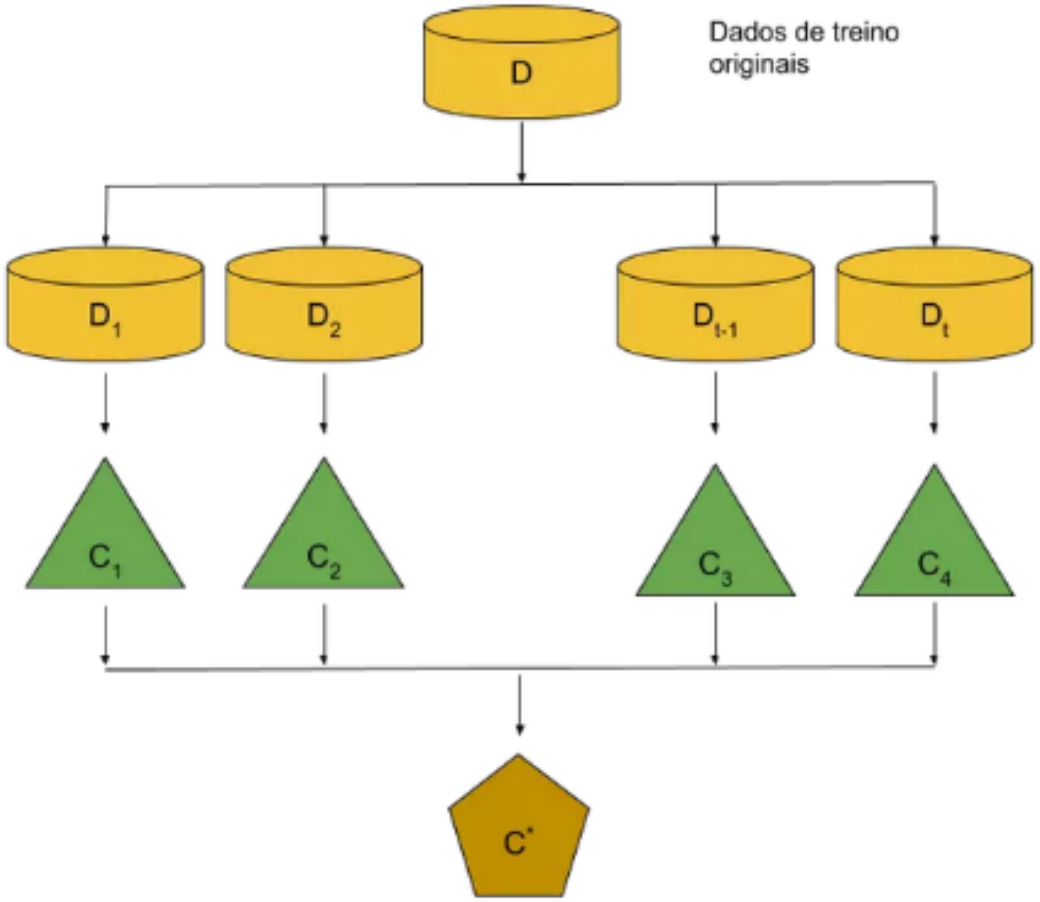
- Alguns dos métodos de ‘ensemble’ comumente utilizados incluem: **‘*Bagging*’, ‘*Boosting*’ e ‘*Stacking*’.**

- ‘***B****ootstrap* ***Agg****regat****ing***’ é uma técnica usada para reduzir a variância das previsões.

- Ela combina o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados.

Métodos de Ensemble

Bagging



Passo 1: criar múltiplos conjuntos de dados

Passo 2: construir múltiplos classificadores

Passo 3: combinar os classificadores

Bootstrap sample ⇛ $f_1(x)$
Bootstrap sample ⇛ $f_2(x)$
Bootstrap sample ⇛ $f_3(x)$
…
Bootstrap sample ⇛ $f_M(x)$

Combine $f_1(x),\ldots, f_M(x)$ ⇛ $f(x)$

MODEL AVERAGING

$f_i(x)$'s are "base learners"

Métodos de Ensemble

Etapas do *Bagging*

- 1º - Criar vários conjuntos de dados:

- A amostragem é feita com a substituição dos dados originais e a formação de novos conjuntos de dados
- Os novos conjuntos de dados podem ter uma fração das colunas e das linhas, que geralmente são hiper-parâmetros em um modelo de '*bagging*'
- Tomando frações de linha e coluna menores que 1 ajuda na montagem de modelos robustos, menos propensos a sobreajuste.

- 2º - Criar múltiplos classificadores

- Classificadores são construídos em cada conjunto de dados
- Em geral, o mesmo classificador é modelado em cada conjunto de dados, e a partir disso as previsões são feitas

Métodos de Ensemble

Etapas do *Bagging*

- 3º - Combinar classificadores

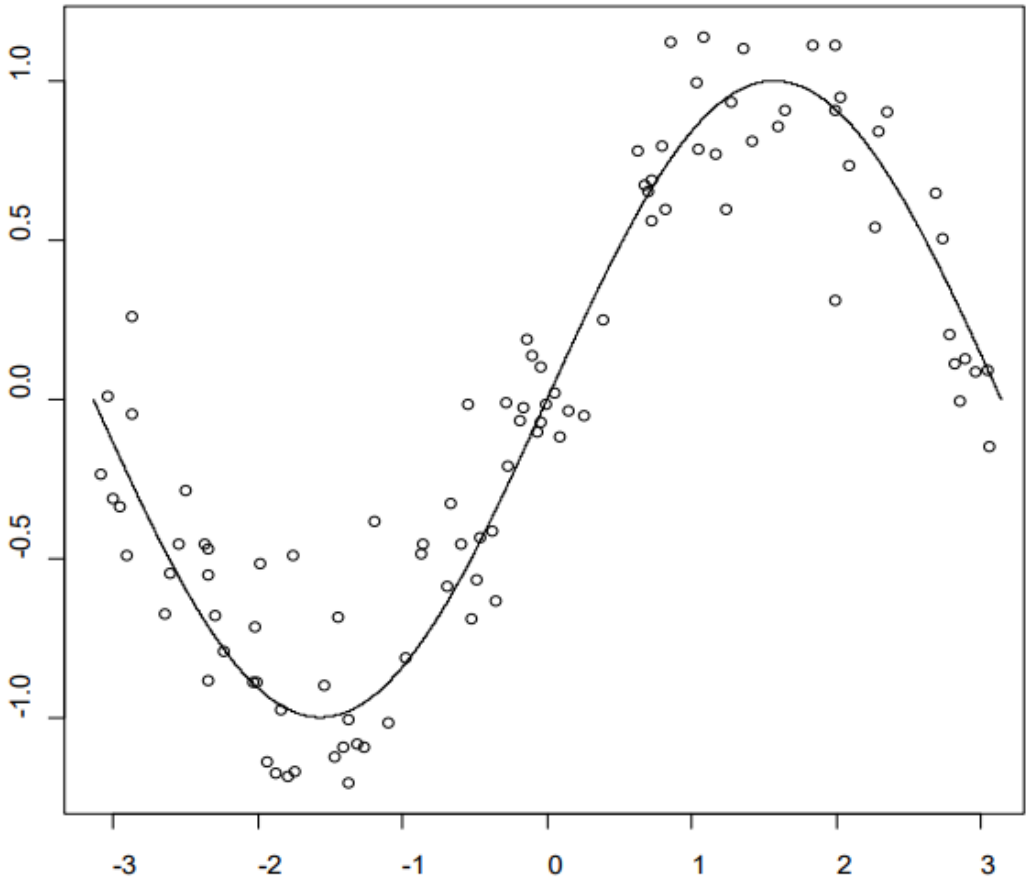
- As previsões de todos os classificadores são combinadas usando a média, a mediana ou a moda, dependendo do problema
- Os valores combinados são geralmente mais robustos do que um único modelo

- Note-se que o número de modelos construídos não são hiper-parâmetros.
- Um maior número de modelos são geralmente *melhores*, **ou** podem dar um *desempenho semelhante* ao de números mais baixos.

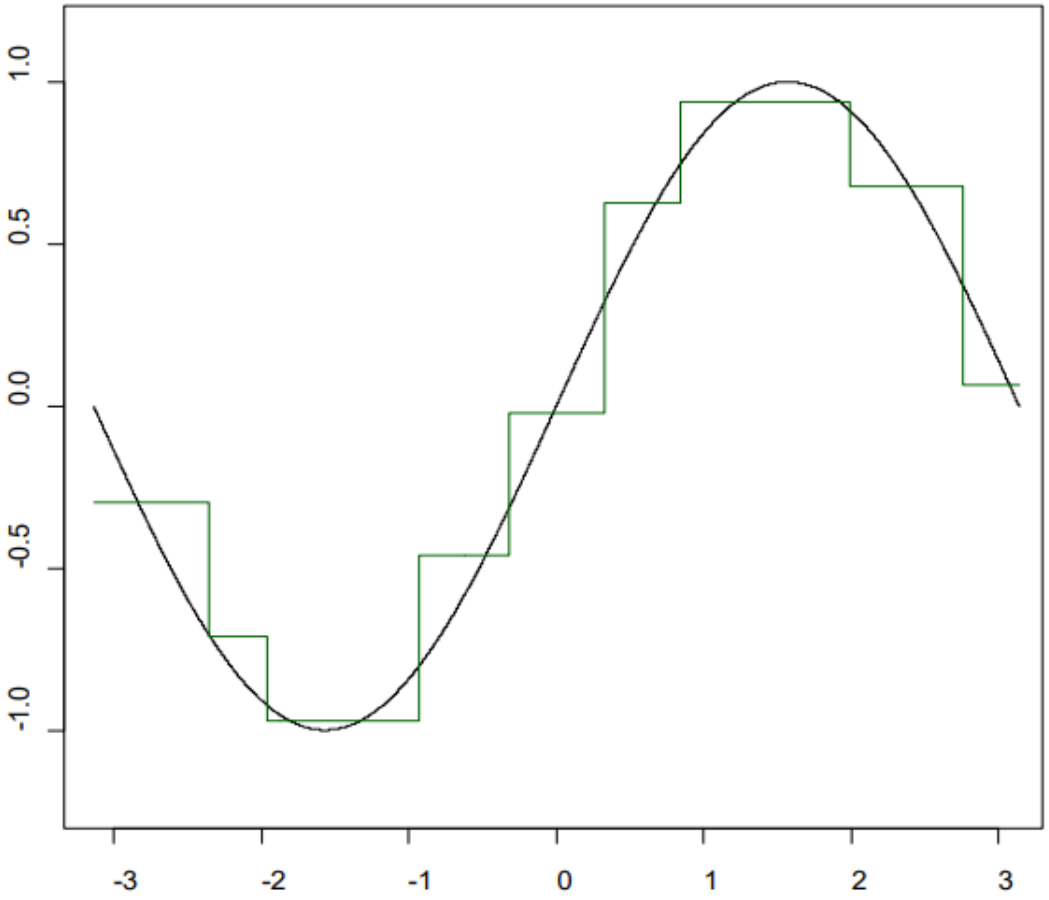
- Existem várias implementações de modelos '*bagging*'. A floresta aleatória ('***random forest***') é uma delas!

Métodos de Ensemble

Exemplo 1



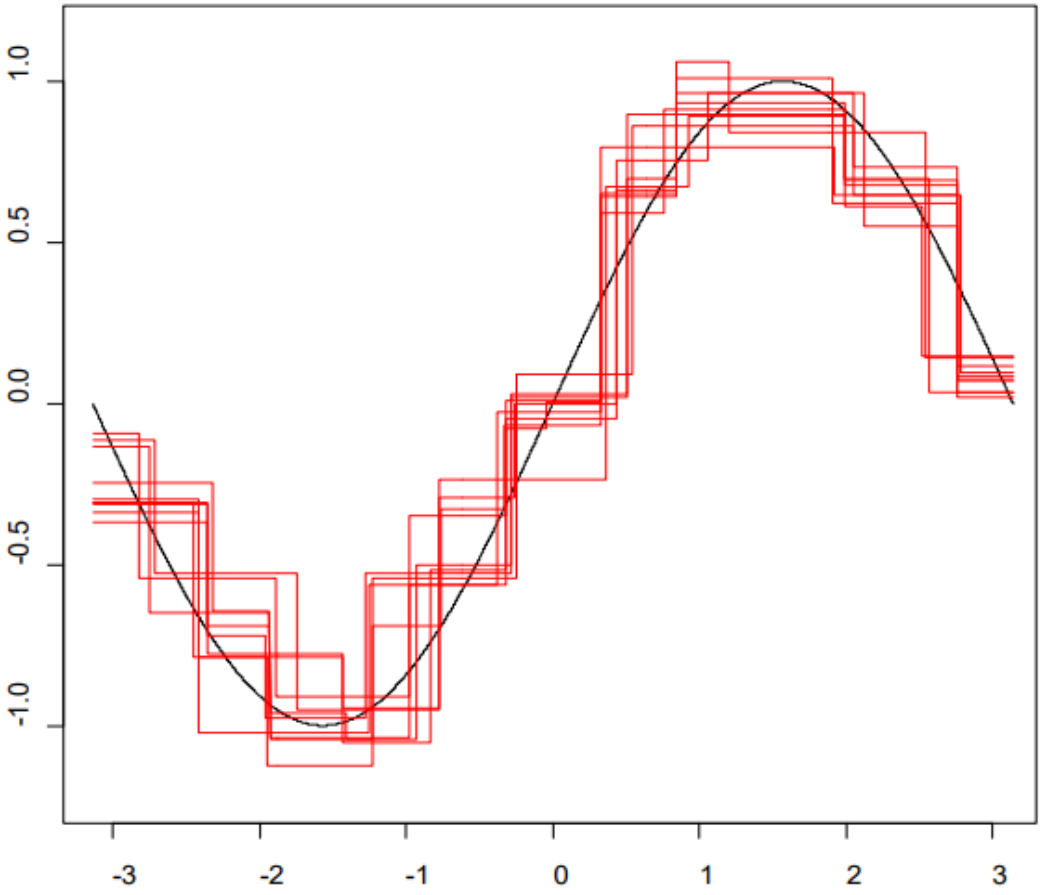
Dados e curva de regressão ajustada.



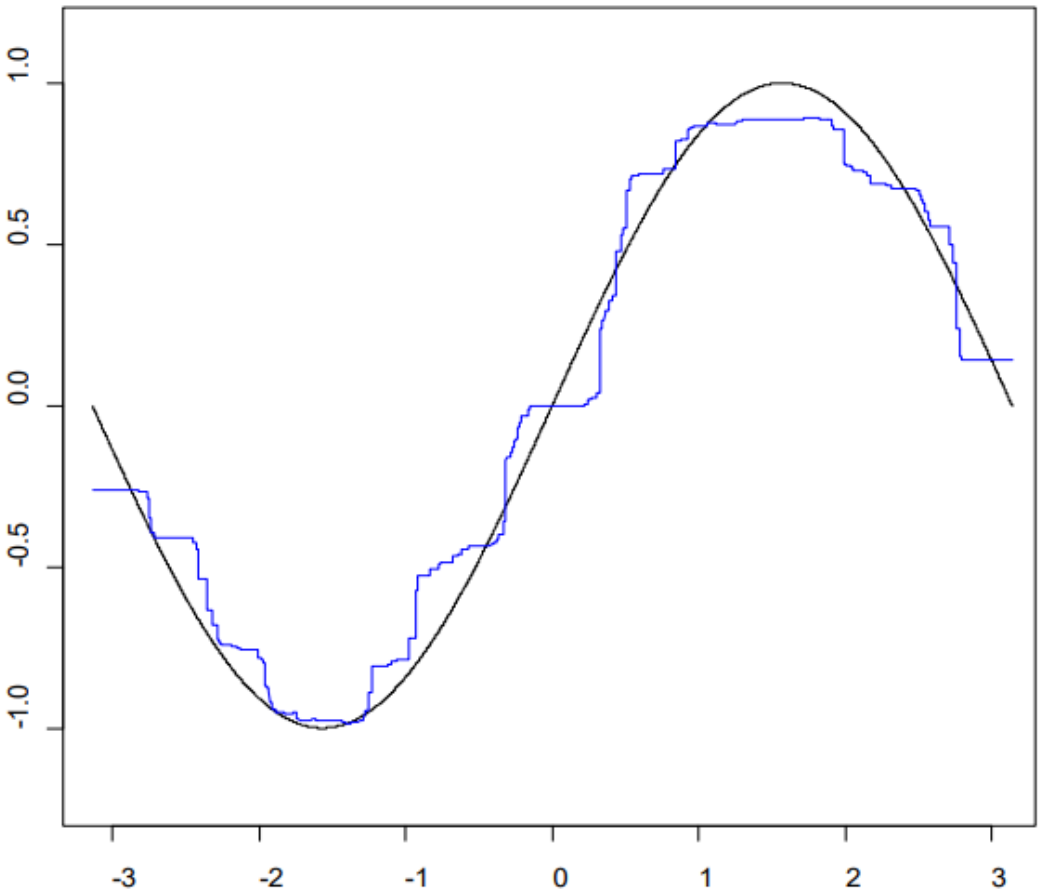
Uma árvore de decisão

Métodos de Ensemble

Exemplo 1



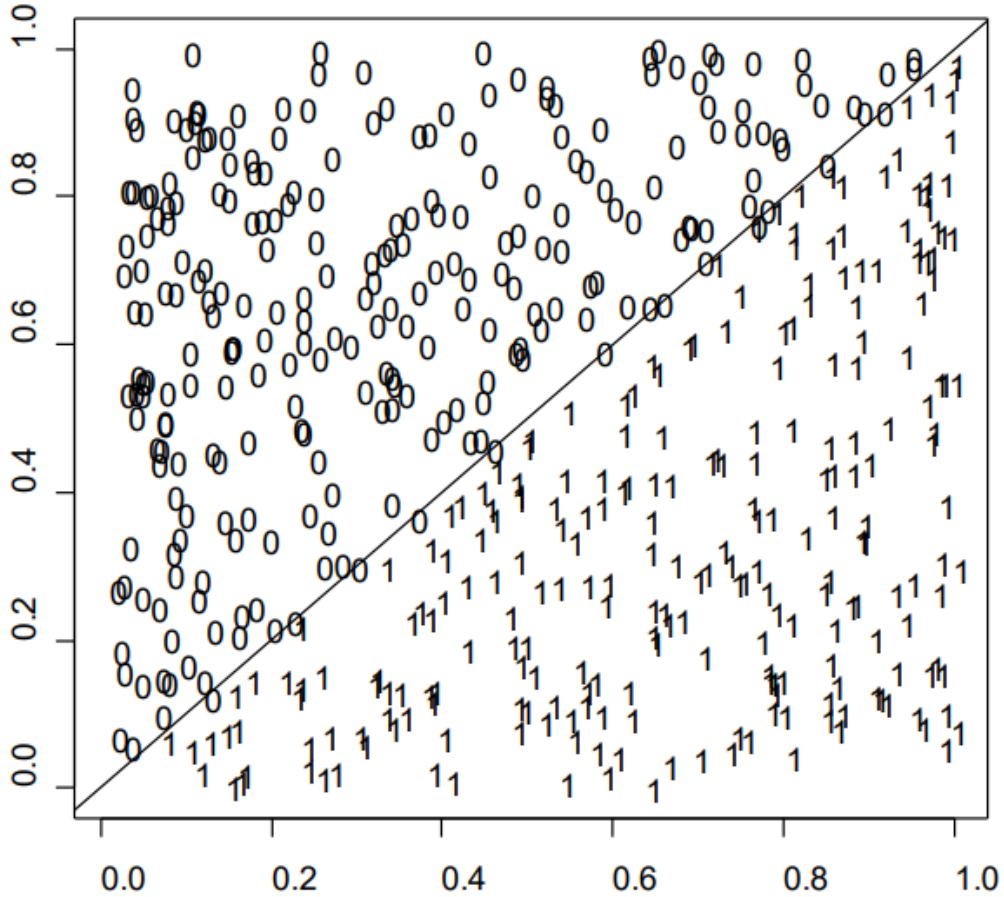
10 árvores de decisão



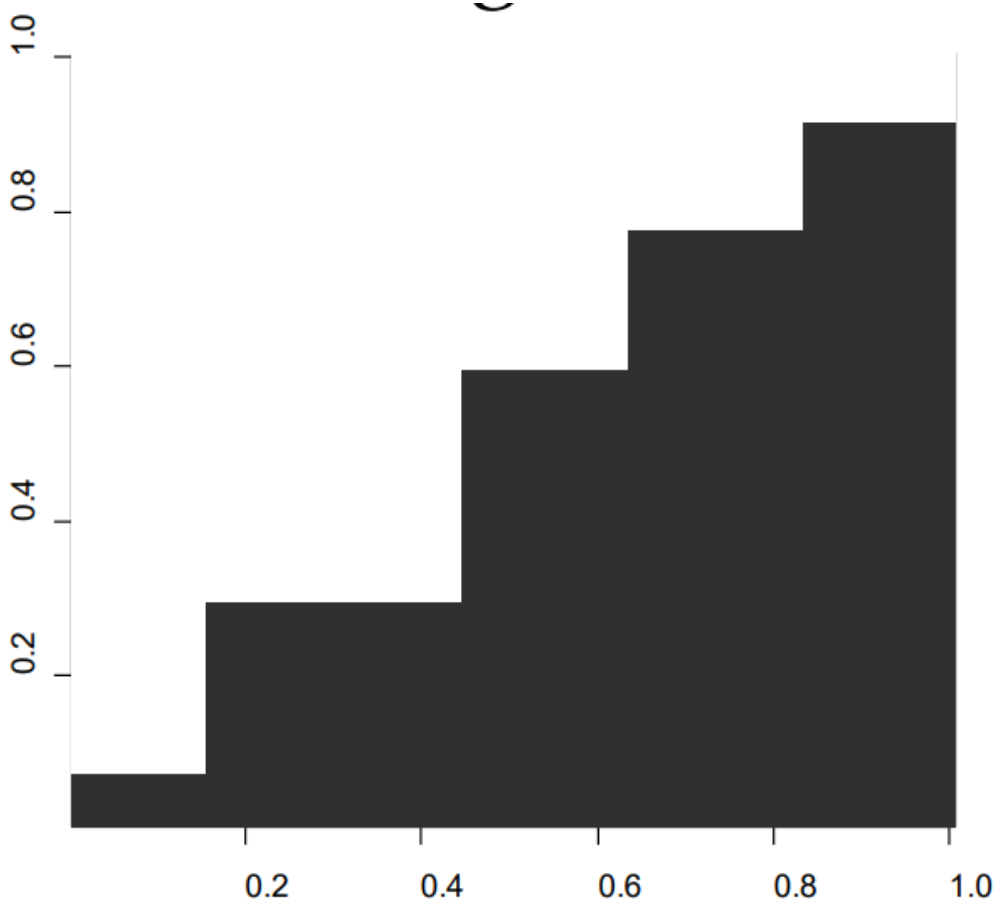
Média de 100 árvores de decisão

Métodos de Ensemble

Exemplo 2



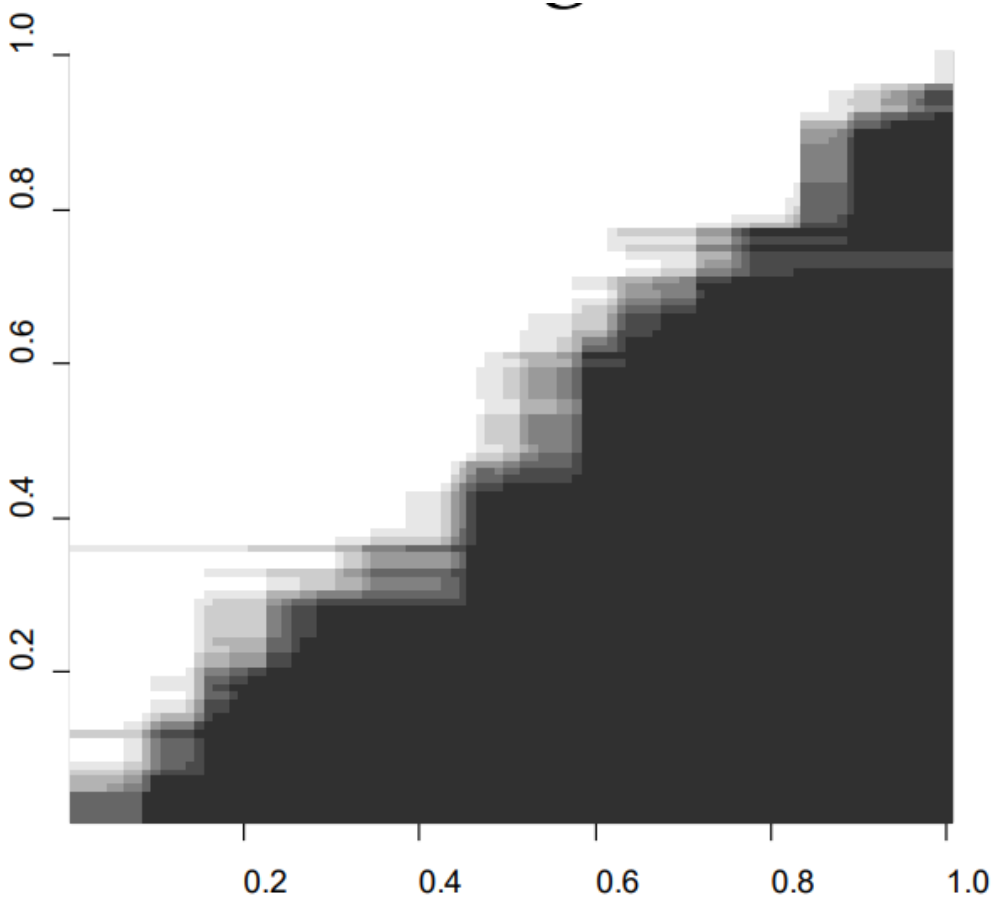
Difícil problema para solução



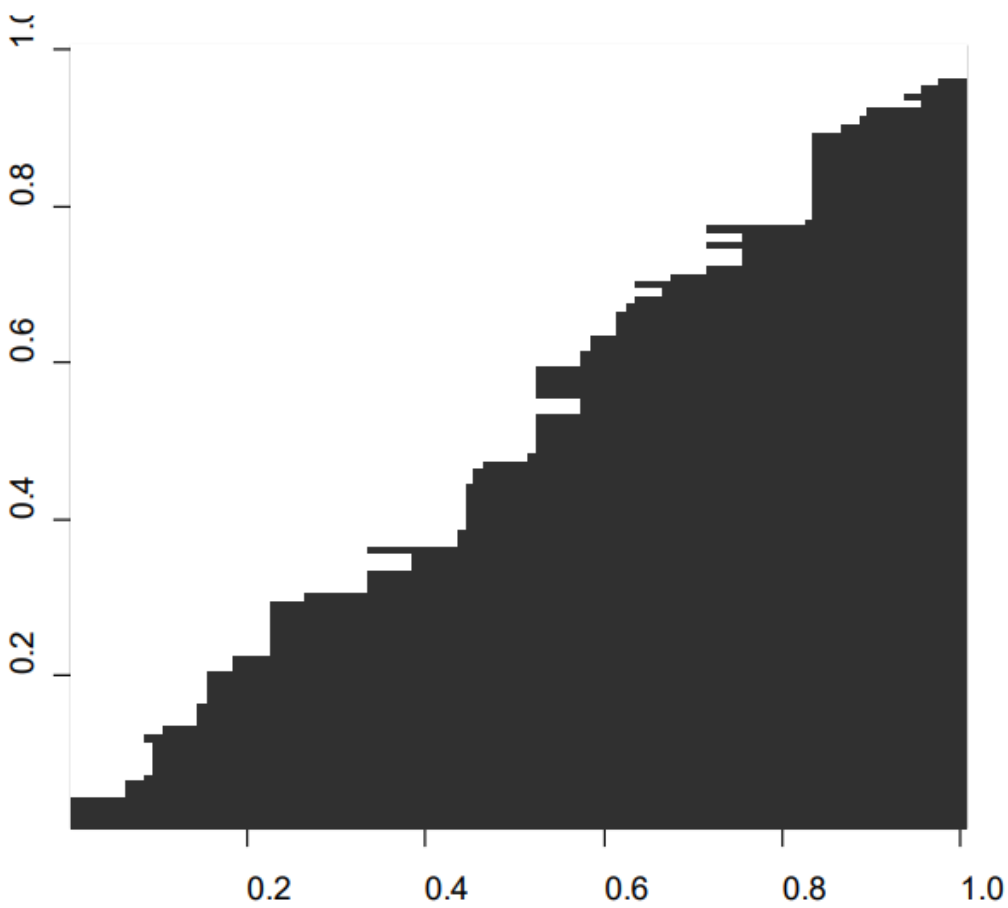
Uma árvore de decisão

Métodos de Ensemble

Exemplo 2



Média de 25 árvores de decisão



Combinação de 25 árvores de decisão

Métodos de Ensemble

Exemplo 3 – A sabedoria das multidões

- Vamos gerar números aleatórios entre 0 – 100.

- Se o número que eu gerar for maior ou igual a 40, você vence (então você tem 60% de chance de vitória) e eu pago a você algum dinheiro. Se for abaixo de 40, eu ganho e você me paga o mesmo valor.

Agora, eu ofereço a você as seguintes opções:

- **Jogo 1 -** jogue 100 vezes, apostando R$ 1 cada vez.
- **Jogo 2 -** jogue 10 vezes, apostando R$ 10 cada vez.
- **Jogo 3 -** jogue uma vez, apostando R$ 100.

Qual você escolheria?

Métodos de Ensemble

Exemplo 3 – A sabedoria das multidões

- Note que, o valor esperado de cada jogo é o mesmo:

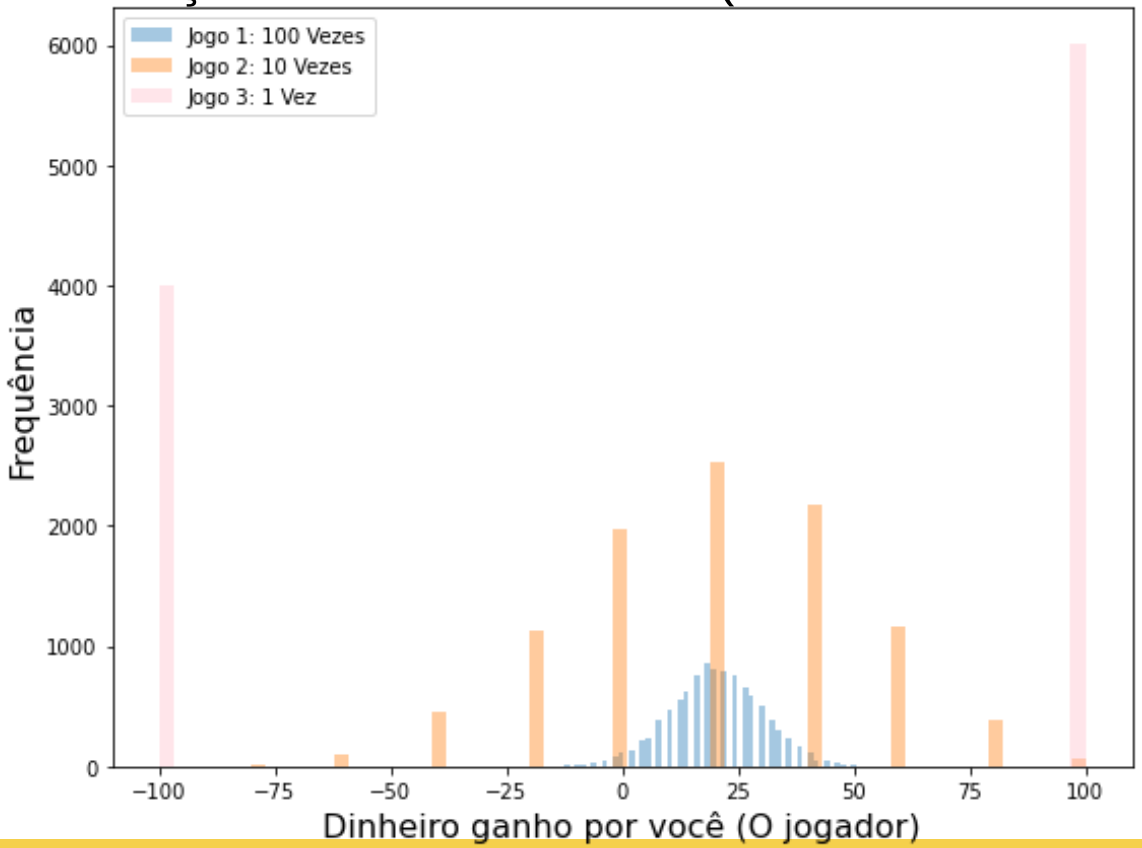
  - **Valor esperado do jogo 1 =** (0,60*1 + 0,40*-1) * 100 **= 20**
  - **Valor esperado do jogo 2 =** (0,60*10 + 0,40*-10) * 10 **= 20**
  - **Valor esperado do jogo 3 = (**0,60*100 + 0,40*-100) * 1 **= 20**

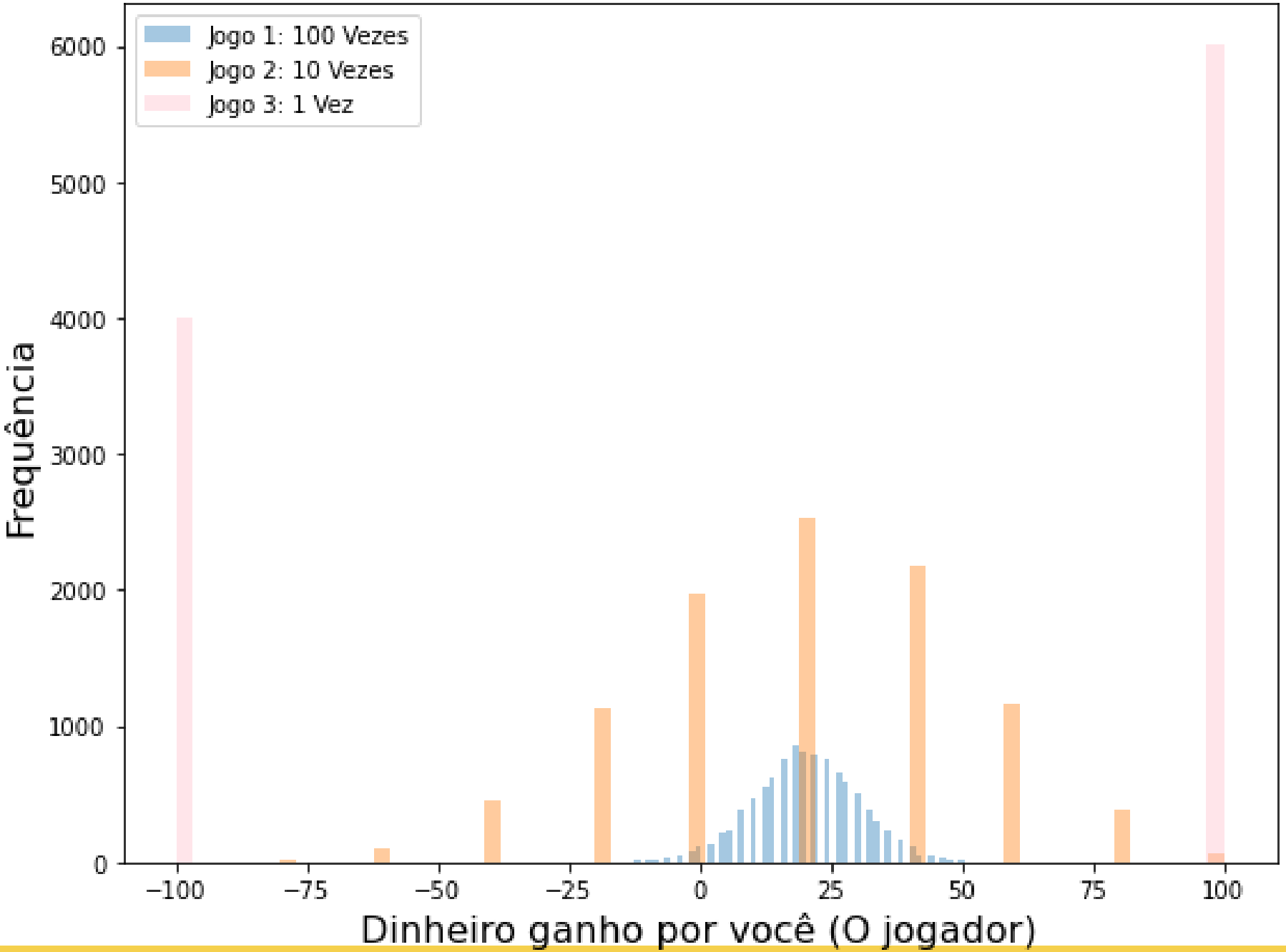
E aí, qual você escolheria?

Métodos de Ensemble

Exemplo 3 – A sabedoria das multidões

- **E quanto às distribuições?**
- Vamos visualizar os resultados com uma simulação de Monte Carlo (executaremos 10.000 simulações de cada tipo de jogo):




Jogo 1: 100 Vezes
Jogo 2: 10 Vezes
Jogo 3: 1 Vez
Frequência
6000
5000
4000
3000
2000
1000
0
−100
−75
−50
−25
0
25
50
75
100
Dinheiro ganho por você (O jogador)

Métodos de Ensemble

Exemplo 3 – A sabedoria das multidões

- Média do Jogo 1: 20,11
- **Probabilidade de ganho no Jogo 1: 0,97 - das 10.000 simulações que fizemos, você ganha dinheiro em 97% delas!**

- Média do Jogo 2: 19,61
- Probabilidade de ganho no Jogo 2: 0,63

- Média do Jogo 3: 20.12
- Probabilidade de ganho no Jogo 3: 0,6

Métodos de Ensemble

Exemplo 3 – A sabedoria das multidões

- Portanto, embora os jogos compartilhem o mesmo valor esperado, suas distribuições de resultados são completamente diferentes.

- Quanto mais dividirmos nossa aposta de R$ 100 em jogadas diferentes, mais confiantes podemos ter de que ganharemos dinheiro.

- **Por quê isso funciona?**

Porque cada peça é independente das outras.

Reconhecimento de padrões e aprendizagem computacional

Random Forest

Leo Breiman, 1928 – 2005

- ✓ 1954: PhD Berkeley (mathematics)
- ✓ 1960 -1967: UCLA (mathematics)
- ✓ 1969 -1982: Consultant
- ✓ 1982 – 1993: Berkeley (statistics)
- ✓ 1984: “Classification & Regression Trees” (with Friedman, Olshen, Stone)
- ✓ 1996: “Bagging”
- ✓ 2001: “Random Forests”

Adele Cutler

- ✓ 1988: PhD Berkeley (mathematics)
- ✓ Orientador: Leo Breiman

  Optimization Methods in Statistics (1988)
- ✓ Professora titular da Utah State University

Random Forest

Definição

Quais palavras devemos colocar aqui?

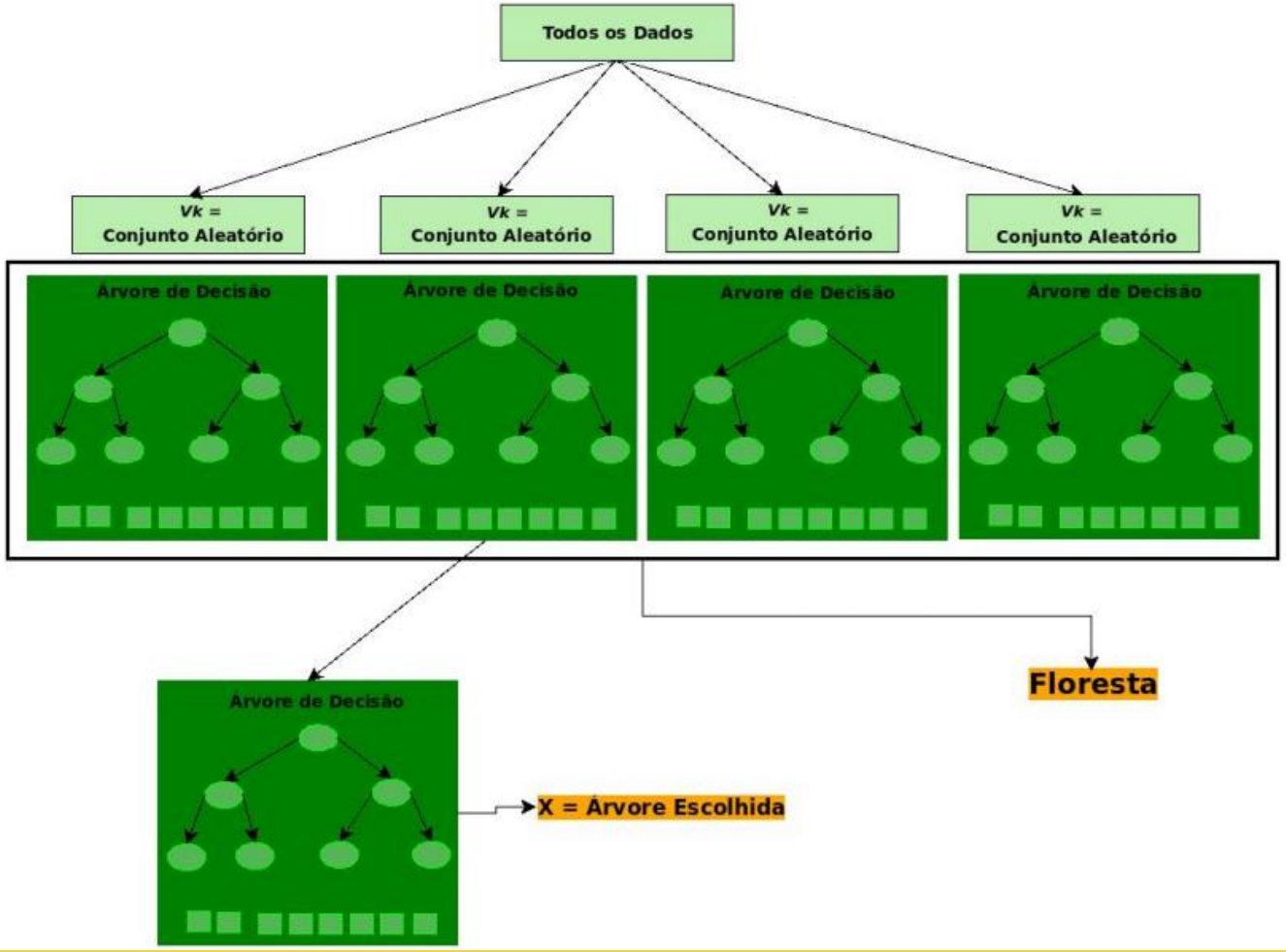
- -

Forme sua definição agora:

“A floresta aleatória é um algoritmo de aprendizado de máquina comumente usado, registrado por *Leo Breiman* e *Adele Cutler*, que combina a saída de várias árvores de decisão para chegar a um único resultado. Sua facilidade de uso e flexibilidade impulsionaram sua adoção, uma vez que lida com problemas de classificação e regressão.”

Random Forest

Funcionamento



Random Forest

Definição

- Floresta Aleatória é considerada a **panacéia** de todos os problemas de **Data Science**. Em outras palavras, quando você não consegue pensar num algoritmo *(seja qual for a situação)*, use ***'random forest'***!

- Ele também aplica métodos de redução dimensional, trata valores faltantes, valores anómalos ('*outliers*') e outras etapas essenciais da **exploração de dados**.

- No geral, faz um trabalho muito bom. É um tipo de método de aprendizado de '*ensemble*', onde um grupo de modelos fracos são combinados para formar um modelo mais forte.

Random Forest

Funcionamento

- 1. Assuma que o número de casos no conjunto de treinamento é N. Então, a amostra desses N casos é escolhida aleatoriamente, mas com substituição. Esta amostra será o conjunto de treinamento para o cultivo da árvore.

- 2. Se houver M variáveis de entrada, um número m << M é especificado de modo que, em cada nó, m variáveis de M sejam selecionadas aleatoriamente. A melhor divisão nestes m é usada para dividir o nó. O valor de m é mantido constante enquanto crescemos a floresta.

- 3. Cada árvore é cultivada na maior extensão possível e não há poda.

- 4. Preveja novos dados agregando as previsões das árvores ntree (ou seja, votos majoritários para classificação, média para regressão).

Random Forest

Exemplo

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
1	Sol	Alta	Alta	Não	Não ir
2	Sol	Alta	Alta	Sim	Não ir
3	Nublado	Alta	Alta	Não	Ir ao campo
4	Chuva	Alta	Alta	Não	Ir ao campo
5	Chuva	Baixa	Normal	Não	Ir ao campo
6	Chuva	Baixa	Normal	Sim	Não ir
7	Nublado	Baixa	Normal	Sim	Ir ao campo
8	Sol	Suave	Alta	Não	Não ir
9	Sol	Baixa	Normal	Não	Ir ao campo
10	Chuva	Suave	Normal	Não	Ir ao campo
11	Sol	Suave	Normal	Sim	Ir ao campo
12	Nublado	Suave	Alta	Sim	Ir ao campo
13	Nublado	Alta	Normal	Não	Ir ao campo
14	Chuva	Suave	Alta	Sim	Não ir

Bootstrap dataset

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
4	Chuva	Alta	Alta	Não	Ir ao campo
11	Sol	Suave	Normal	Sim	Ir ao campo
2	Sol	Alta	Alta	Sim	Não ir
12	Nublado	Suave	Alta	Sim	Ir ao campo

Random Forest

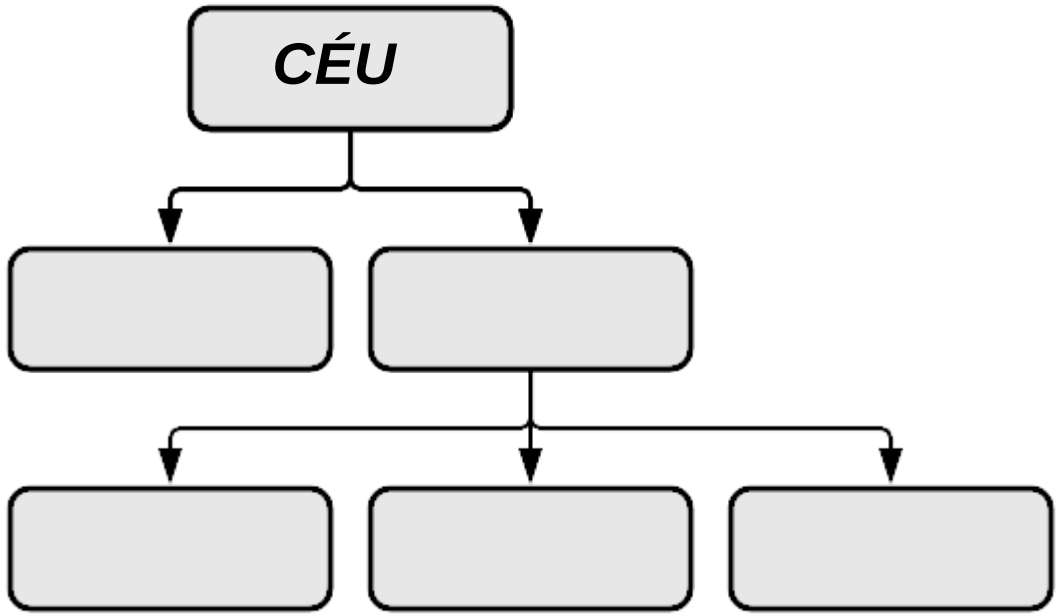
Exemplo

Bootstrap dataset

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
4	Chuva	Alta	Alta	Não	Ir ao campo
11	Sol	Suave	Normal	Sim	Ir ao campo
2	Sol	Alta	Alta	Sim	Não ir
12	Nublado	Suave	Alta	Sim	Ir ao campo

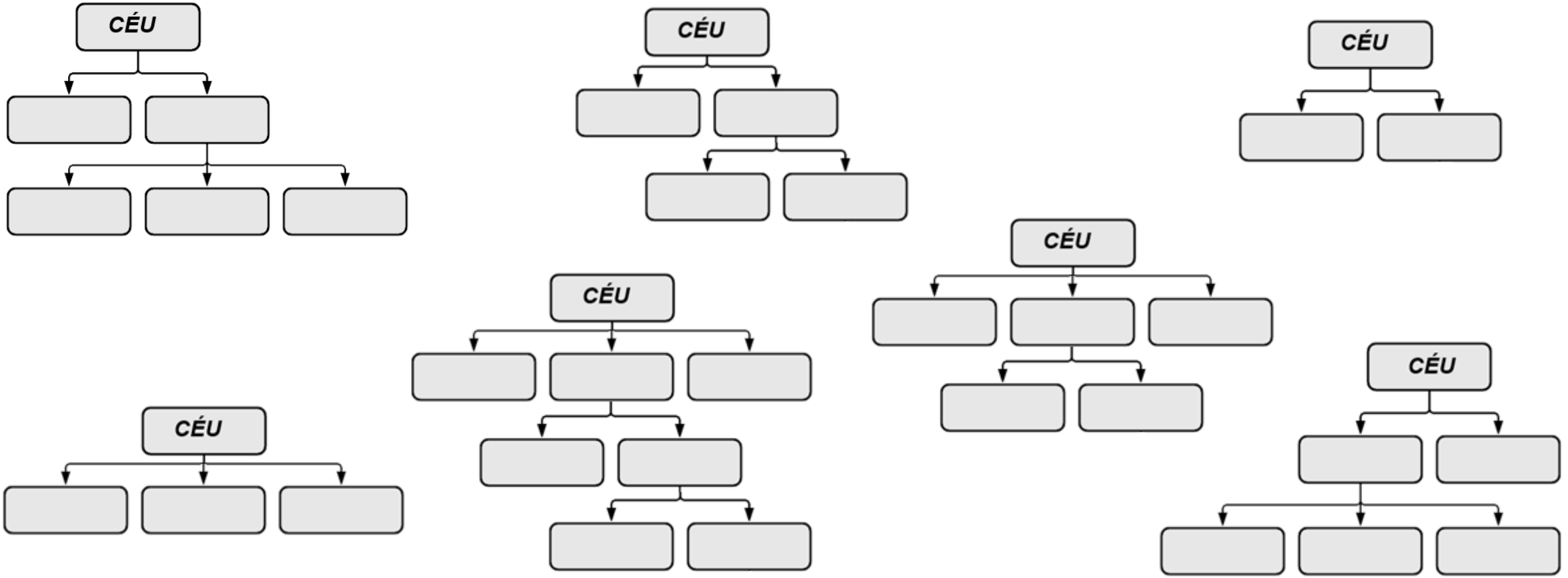
***N features* - aleatória**

CÉU	TEMP.
Chuva	Alta
Sol	Suave
Sol	Alta
Nublado	Suave



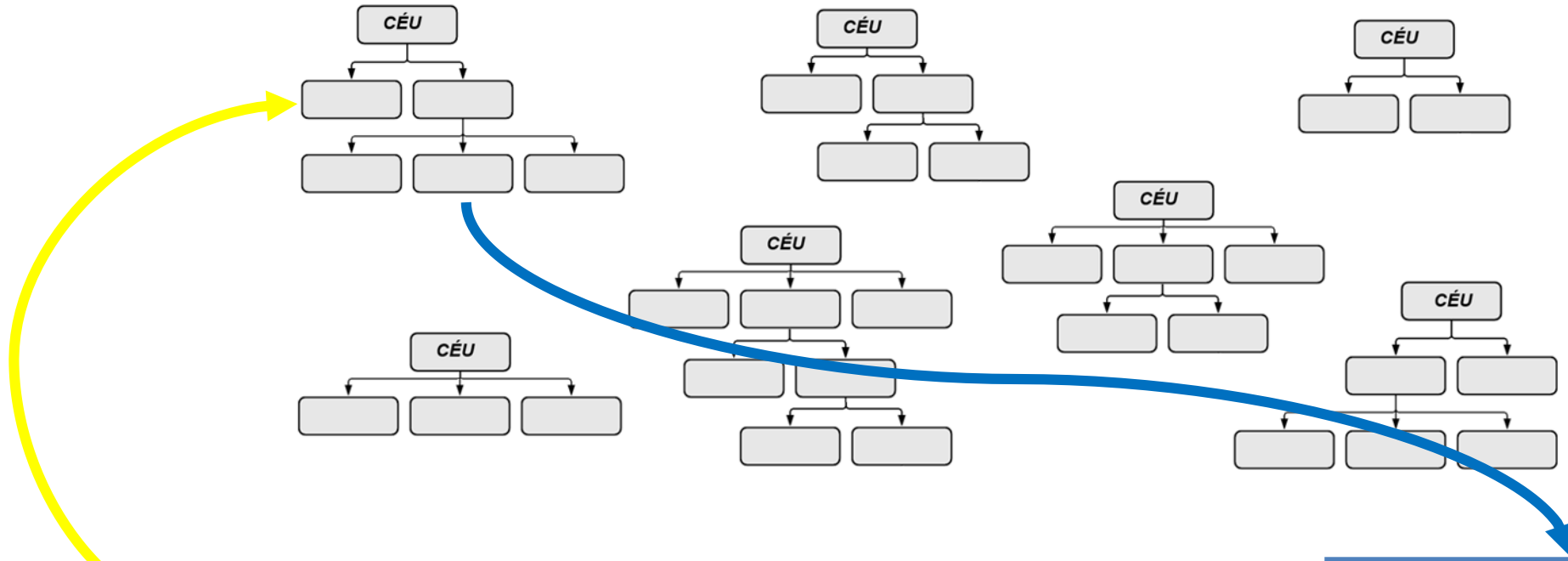
Random Forest

Exemplo



Random Forest

Exemplo



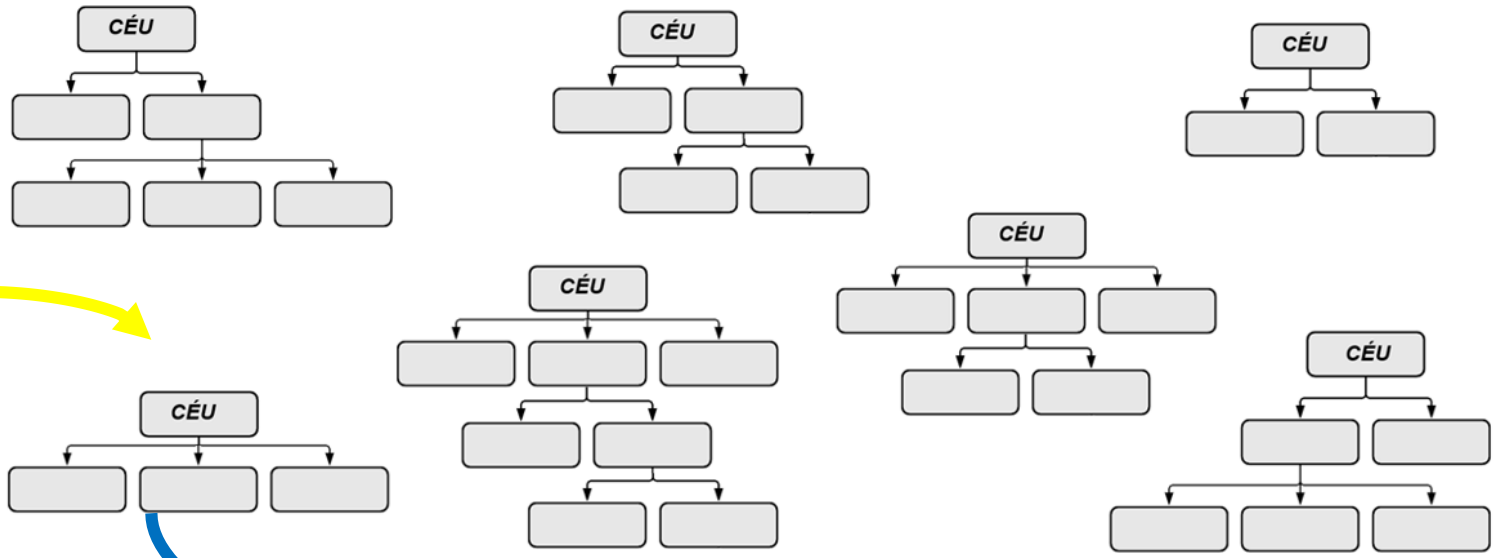
Ir ao campo?	
Sim	Não
0	1

E agora?

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
14	Chuva	Suave	Alta	Sim	??????

Random Forest

Exemplo



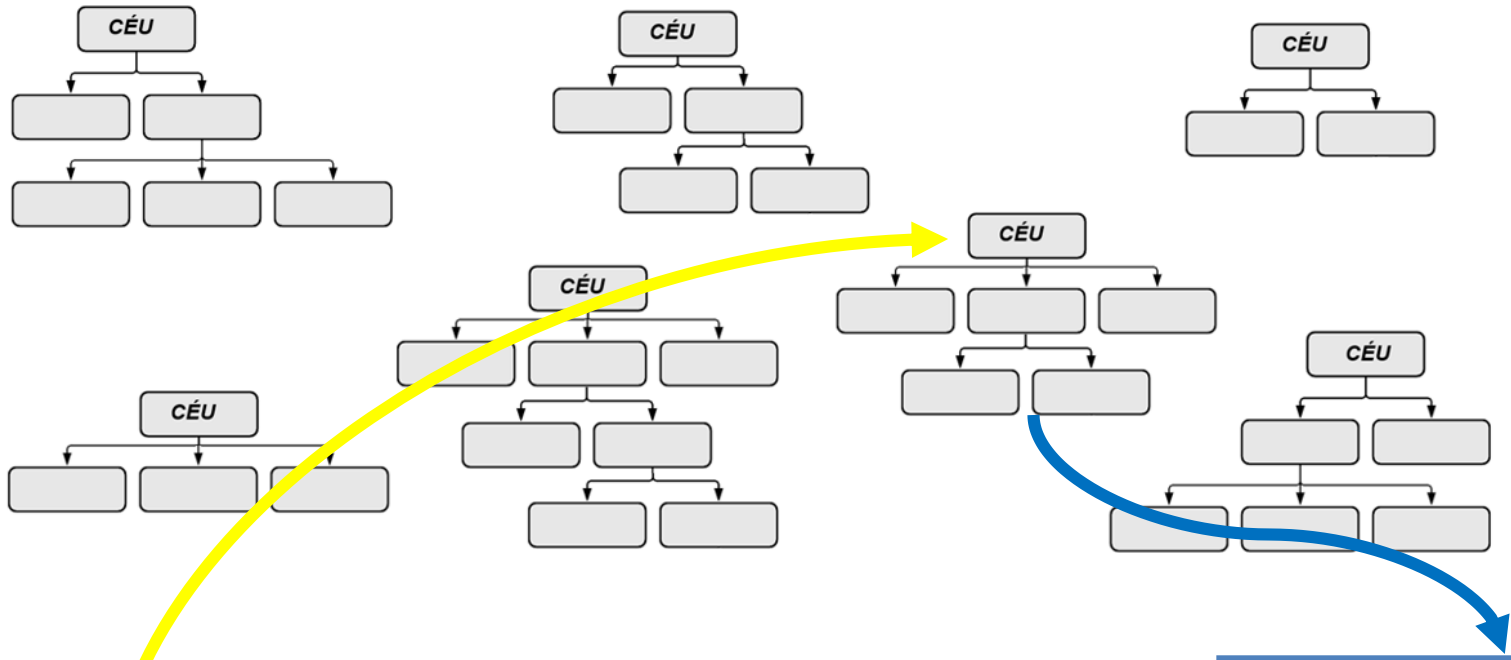
E agora?

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
14	Chuva	Suave	Alta	Sim	??????

Ir ao campo?	
Sim	Não
0	2

Random Forest

Exemplo



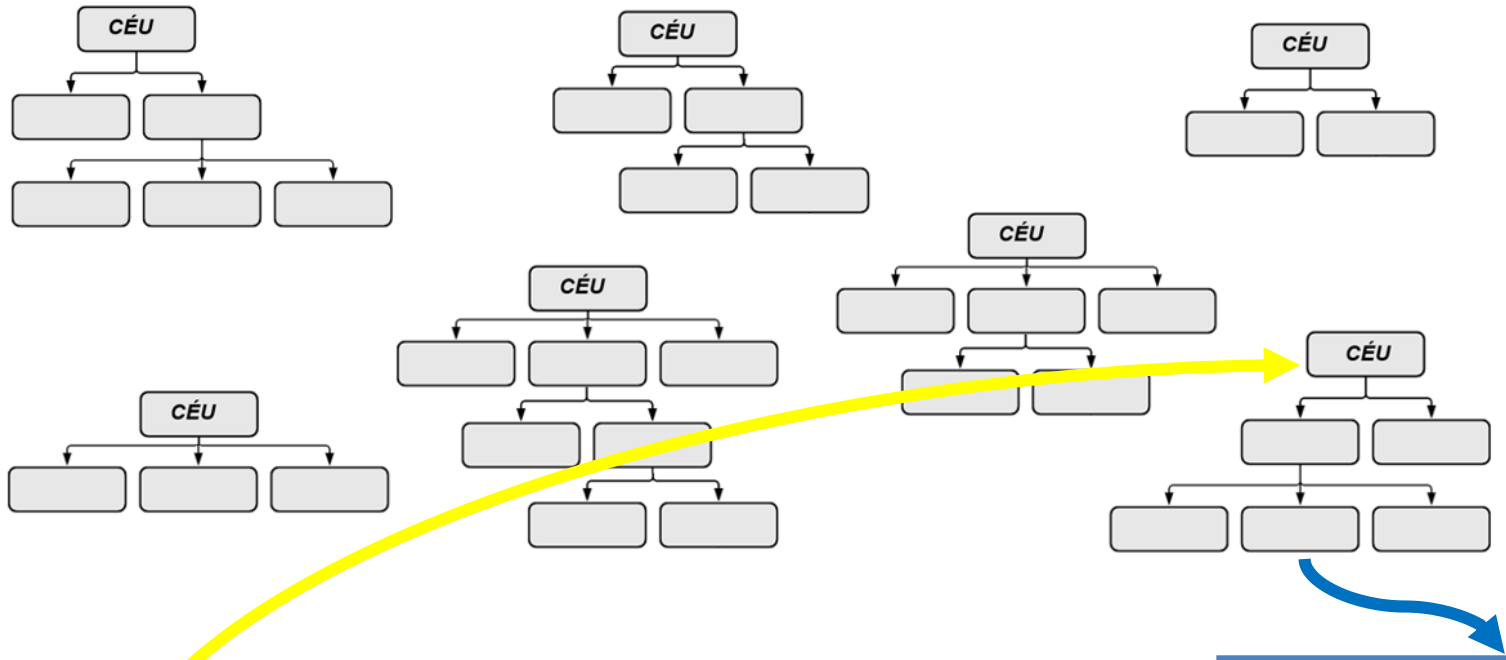
E agora?

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
14	Chuva	Suave	Alta	Sim	??????

Ir ao campo?	
Sim	Não
1	2

Random Forest

Exemplo



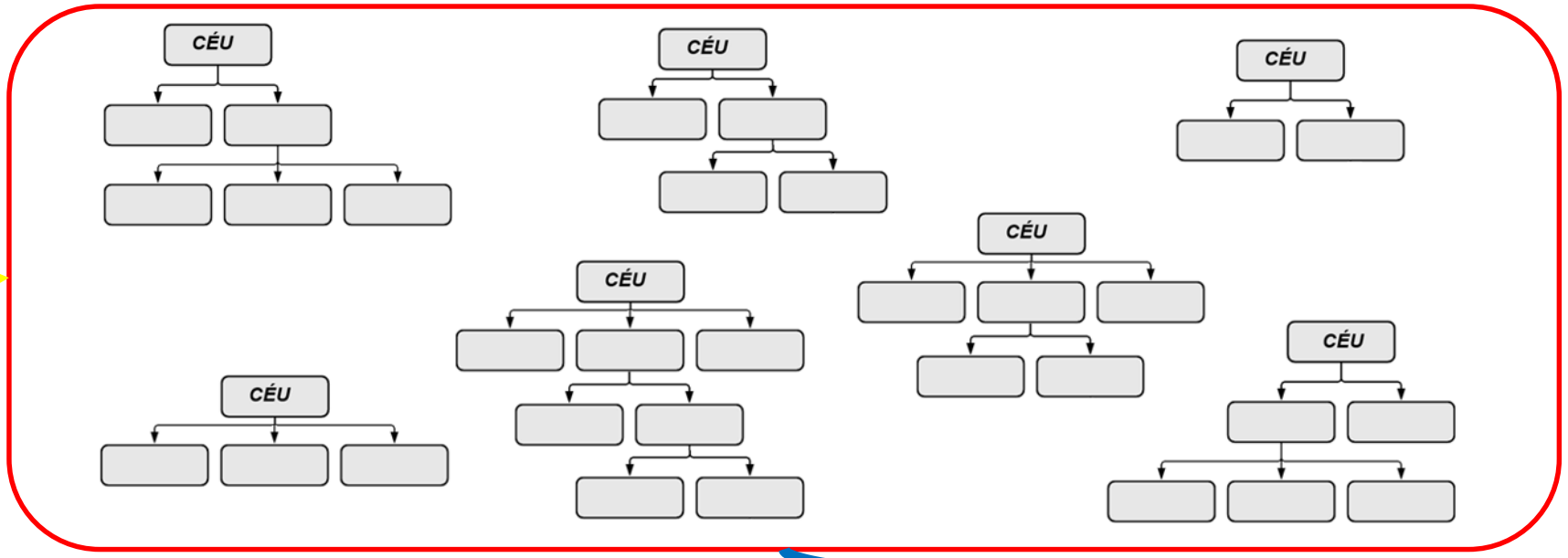
E agora?

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
14	Chuva	Suave	Alta	Sim	??????

Ir ao campo?	
Sim	Não
1	3

Random Forest

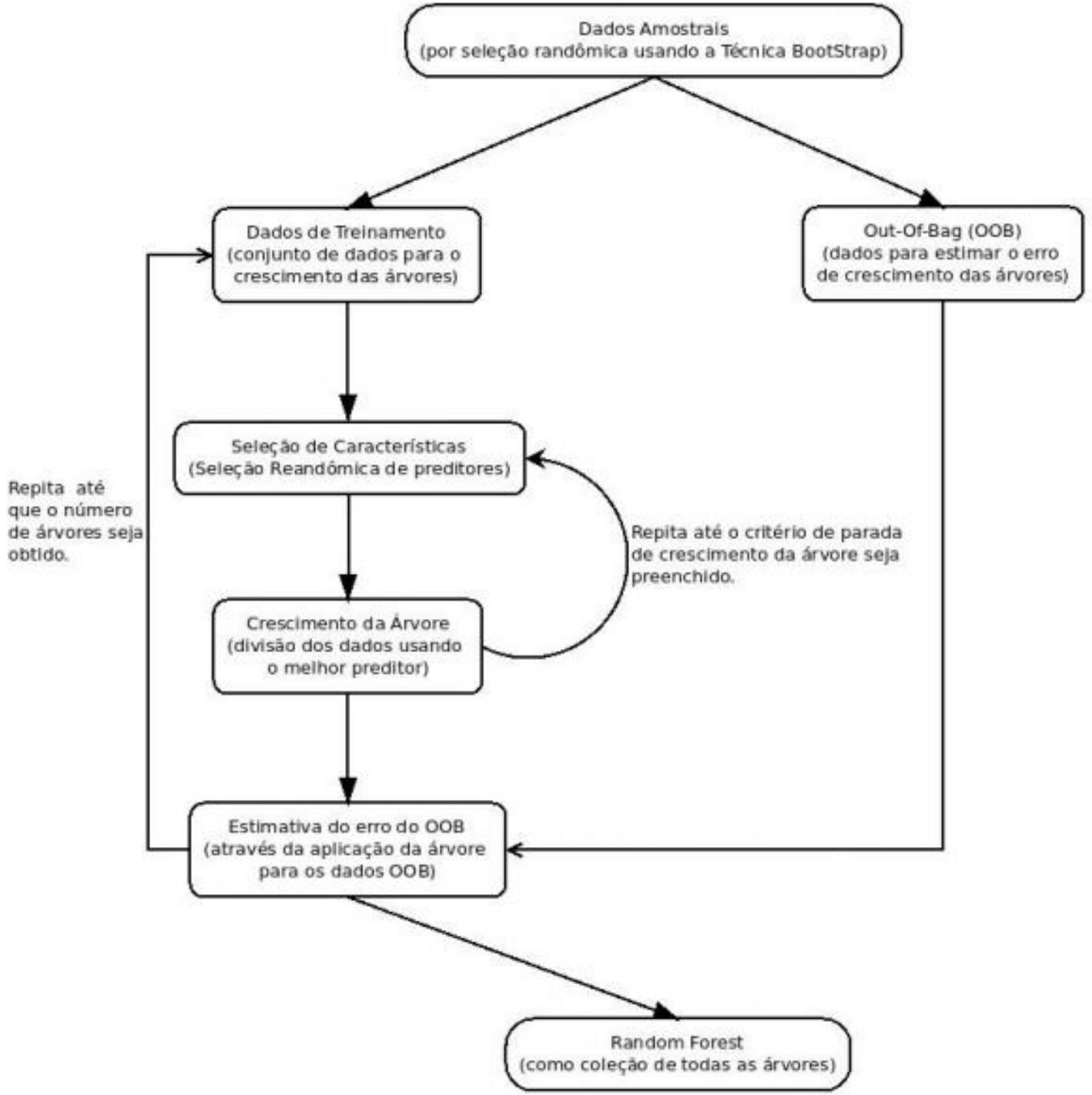
Exemplo



E agora?

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
14	Chuva	Suave	Alta	Sim	**Não ir ao campo**

Ir ao campo?	
Sim	Não
2	**5**


Dados Amostrais
(por seleção randômica usando a Técnica BootStrap)
Dados de Treinamento
(conjunto de dados para o
crescimento das árvores)
Out-Of-Bag (OOB)
(dados para estimar o erro
de crescimento das árvores)
Seleção de Características
(Seleção Reandômica de preditores)
Repita até o critério de parada
de crescimento da árvore seja
preenchido.
Repita até
que o número
de árvores seja
obtido.
Crescimento da Árvore
(divisão dos dados usando
o melhor preditor)
Estimativa do erro do OOB
(através da aplicação da árvore
para os dados OOB)
Random Forest
(como coleção de todas as árvores)

Random Forest

Vantagens

- Este algoritmo pode resolver os problemas de classificação e de regressão, fazendo uma estimativa decente em ambos.

- Um dos benefícios da floresta aleatória que me agrada mais é o poder de lidar com dados em grandes volumes e com muitas dimensões.
  - Ele pode lidar com milhares de variáveis de entrada e identificar as variáveis mais significativas, sendo por isso considerado um dos métodos de redução de dimensões.
  - Além disso, o modelo produz o grau de importância das variáveis, o que pode ser um dado muito útil (em algum conjunto de dados aleatórios).

Random Forest

Vantagens

- Possui um método eficaz para estimar os dados faltantes e mantém a precisão quando uma grande parte dos dados estão faltando.

- Possui métodos para equilibrar erros em conjuntos de dados onde as classes são desequilibradas.

- As capacidades do método anterior podem ser estendidas para dados sem rótulo, levando a clusters não supervisionados, visualizações de dados e detecção de ‘outliers’.

Random Forest

Vantagens

- A floresta aleatória envolve a amostragem dos dados de entrada com substituição chamada como amostragem de ‘bootstrap’.
  - Aqui um terço dos dados não é usado para treinamento e pode ser usado para testes. Estes são chamados de amostras de fora da cesta.
  - O erro estimado nas amostras de fora da cesta é conhecido como erro de fora da cesta.
  - O estudo de estimativas do erro de fora da cesta fornece evidências para mostrar que a estimativa de fora da cesta é tão precisa quanto usar um conjunto de teste do mesmo tamanho que o conjunto de treinamento.
  - Portanto, usar a estimativa de erro de fora da cesta remove a necessidade de ter um conjunto de teste extra.

Random Forest

Desvantagens

- Enquanto faz um bom trabalho na classificação, já não é tão bom para o problema de regressão, uma vez que não fornece previsões precisas para variáveis contínuas.
  - No caso da regressão, não prevê além do intervalo dos dados de treinamento, e que eles podem sobre-ajustar os conjuntos de dados que tenham muita discrepância (‘*noisy*’).

- A floresta aleatória pode ser considerada como uma caixa preta para quem faz modelagem estatística – você tem muito pouco controle sobre o que o modelo faz.
  - Você pode, na melhor das hipóteses, experimentar diferentes parâmetros.


!!! SIMBORA !!!