ENSEMBLE METHODS

GRUPO 3

INTEGRANTES

ANA OLIVEIRA

ANDRÉ SOUZA

BRUNO DE SOUSA DONATO

LUCAS DOS SANTOS GARCIA

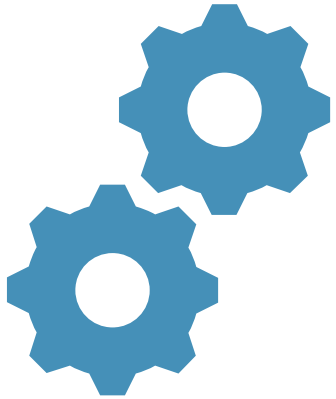
NICOLAS SPOGIS

ROGER TREZZA

THIAGO MARTINS

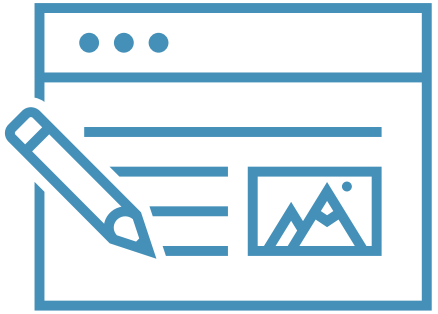
THOMAZ BARROS

VINICIUS VIZENZO



Funcionamento

Principais Métodos



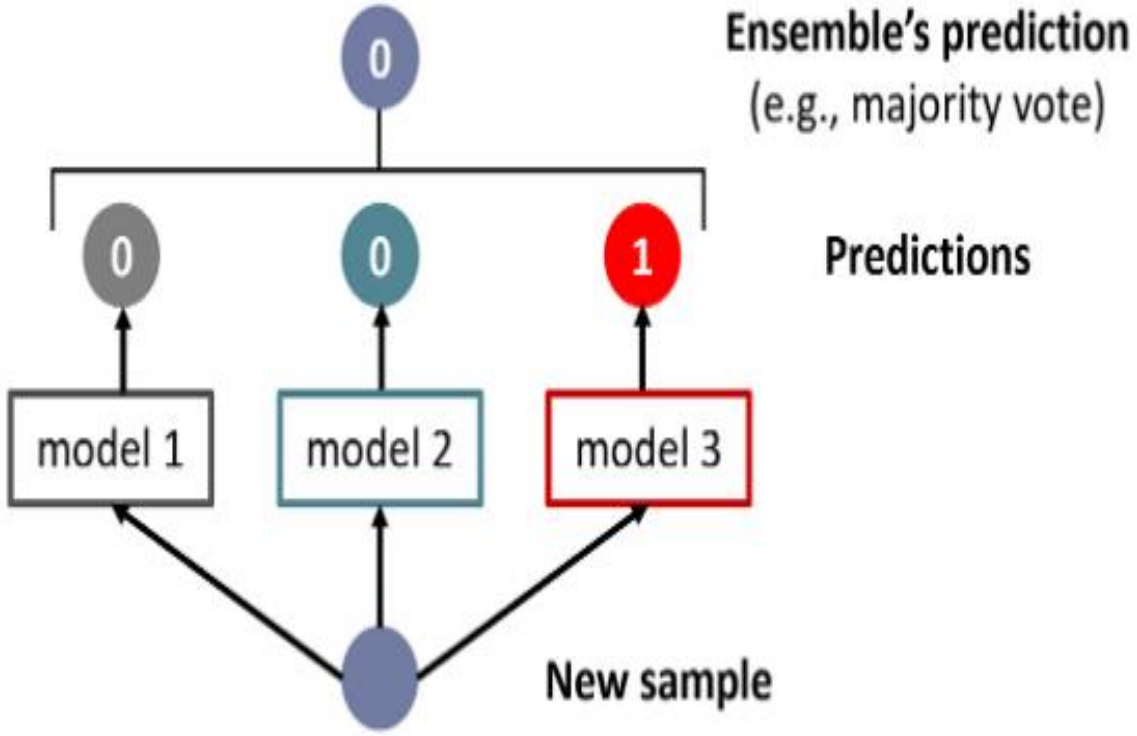
Benefícios e Desvantagens

Ensemble refere-se a uma técnica em aprendizado de máquina onde múltiplos modelos são combinados para melhorar o desempenho geral.

ENSEMBLE METHODS

Ensemble Methods podem ser traduzidos livremente como “Métodos de Conjunto”. Esta tradução faz sentido, pois as técnicas de ensemble consistem justamente em combinar múltiplos modelos individuais para tentar melhorar a performance preditiva sobre um determinado problema.

Podemos fazer uma analogia desta técnica a um princípio conhecido como “A Sabedoria das Multidões”, que estabelece que estimativas e respostas mais precisas podem ser obtidas combinando os julgamentos de diferentes avaliadores.



institut d'intelligence artificielle appliquée

"Muitos são mais inteligentes que alguns, e a inteligência coletiva pode transformar os negócios, a economia, a sociedade e as nações"

JAMES SUROWIECKI:
THE WISDOM OF CROWDS (2004)

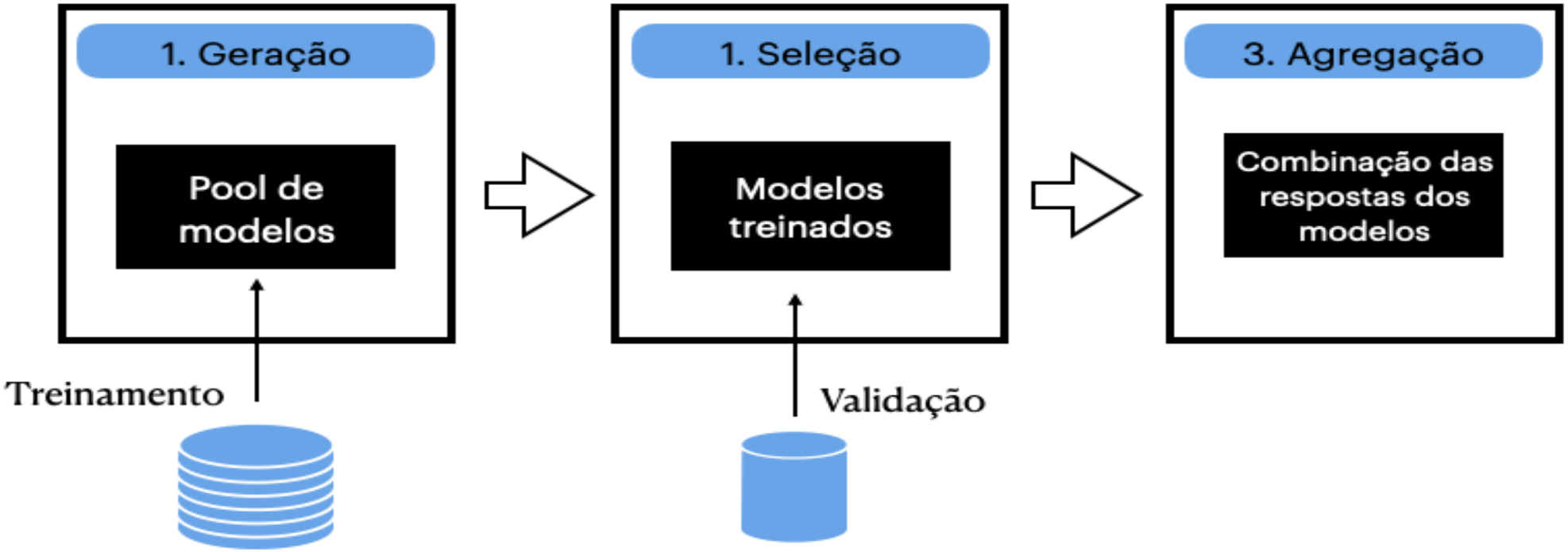
I² A² institut d'intelligence artificielle appliquée

COMBINANDO OS MODELOS

Ao selecionarmos os modelos que iremos combinar, comumente utilizamos um algoritmo de aprendizado para modelos fracos homogêneos (Bagging e Boosting).

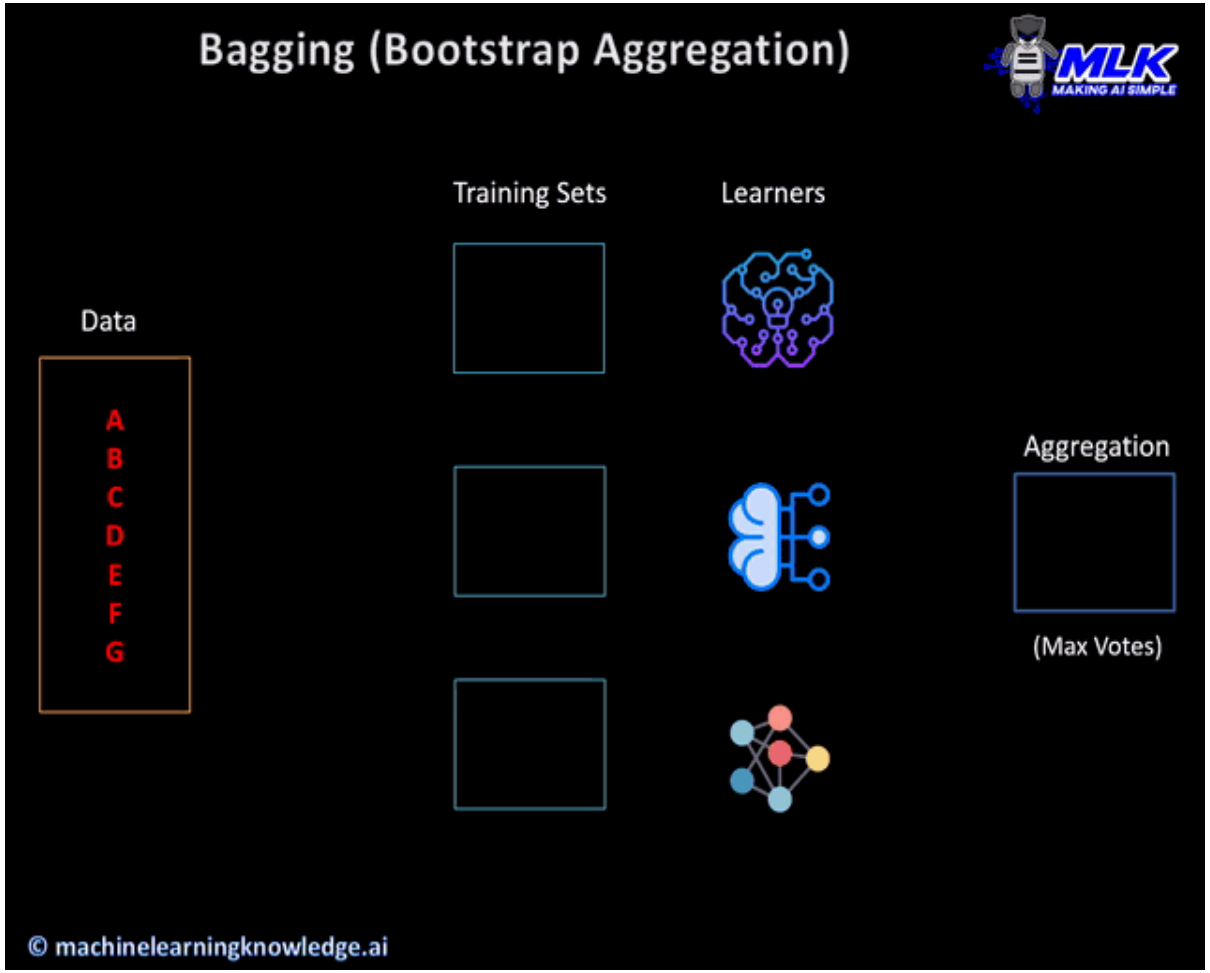
No entanto, existem métodos de ensemble que utilizam algoritmos diferentes para combinar modelos heterogêneos (Stacking).

institut d'intelligence artificielle appliquée



A ideia é que a combinação de vários modelos possa compensar as fraquezas individuais de cada modelo, resultando em uma previsão mais precisa e robusta.

BAGGING



Consiste em reunir vários modelos independentes e encontrar a "média" das previsões com o intuito de obter um modelo de variância.

Boostrapping

institut d'intelligence artificielle appliquée

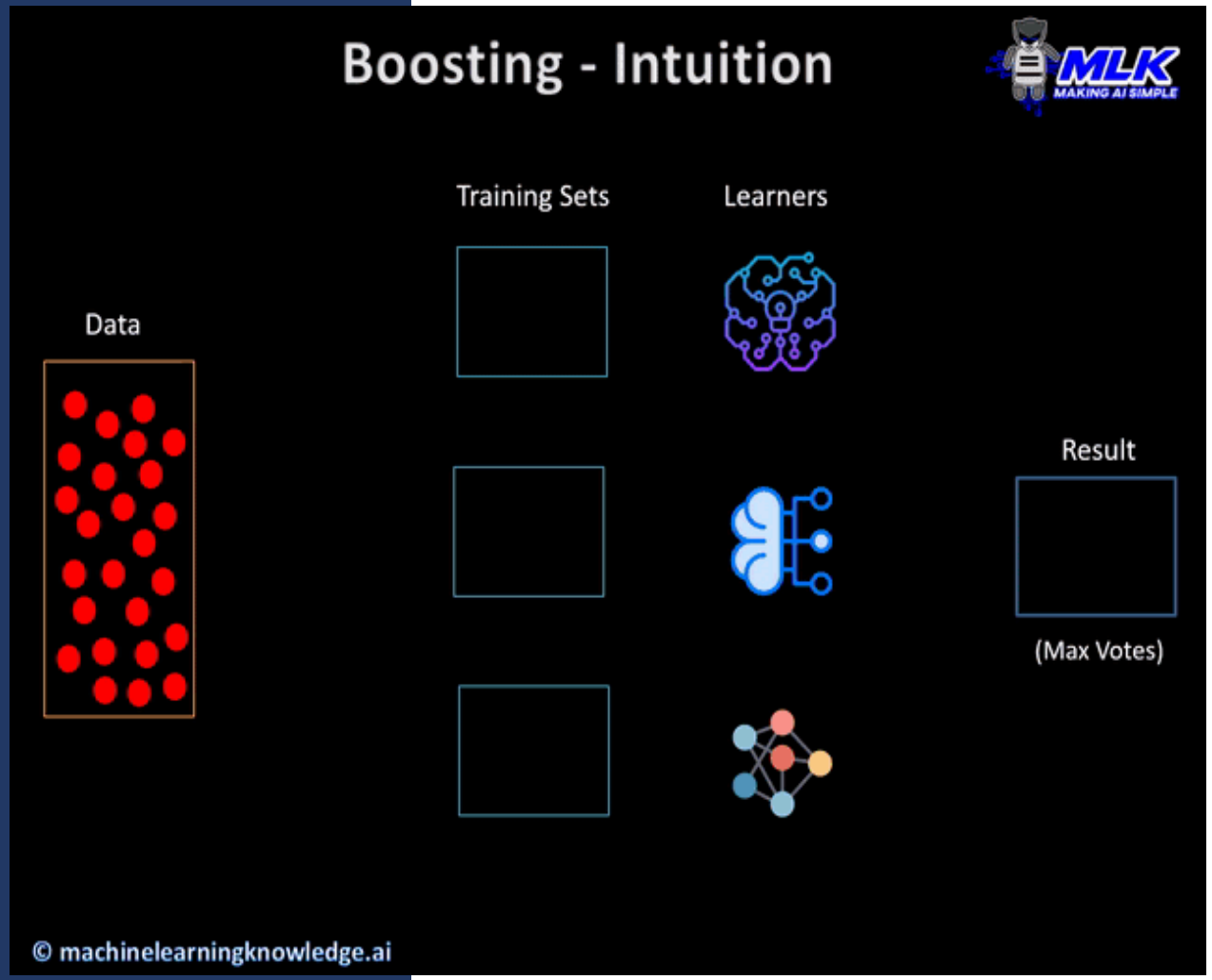
BOOSTING

COMO É FEITO O TREINAMENTO EM BOOSTING?

Etapa 1

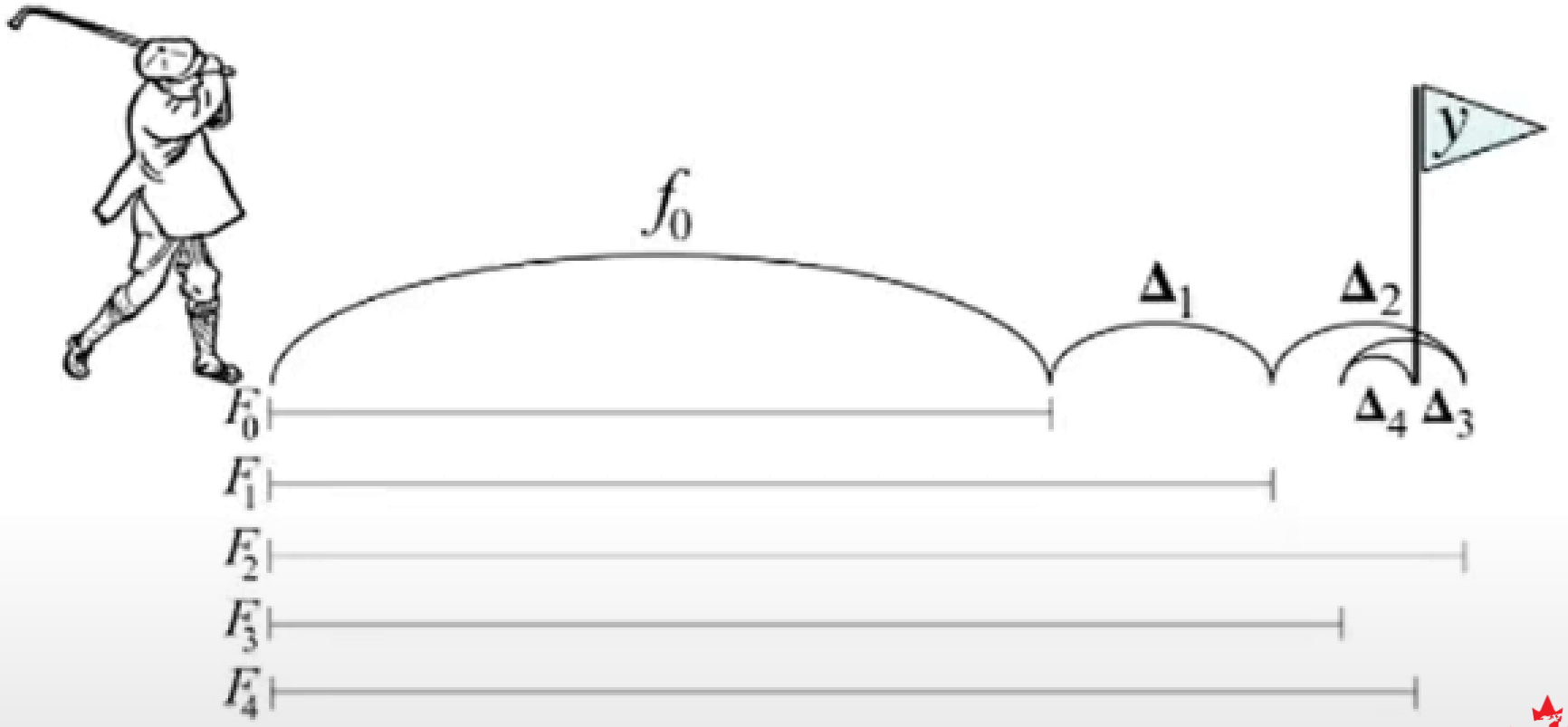
Etapa 2

Etapa 3



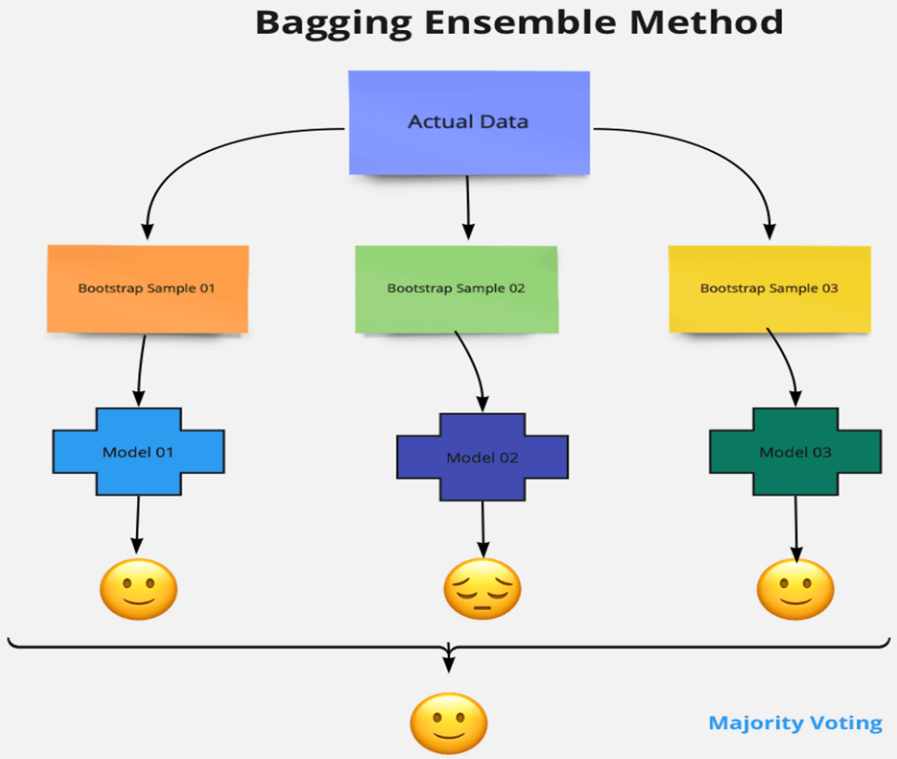
institut d'intelligence artificielle appliquée

A cada passo nosso modelo aprender a ir um pouco melhor

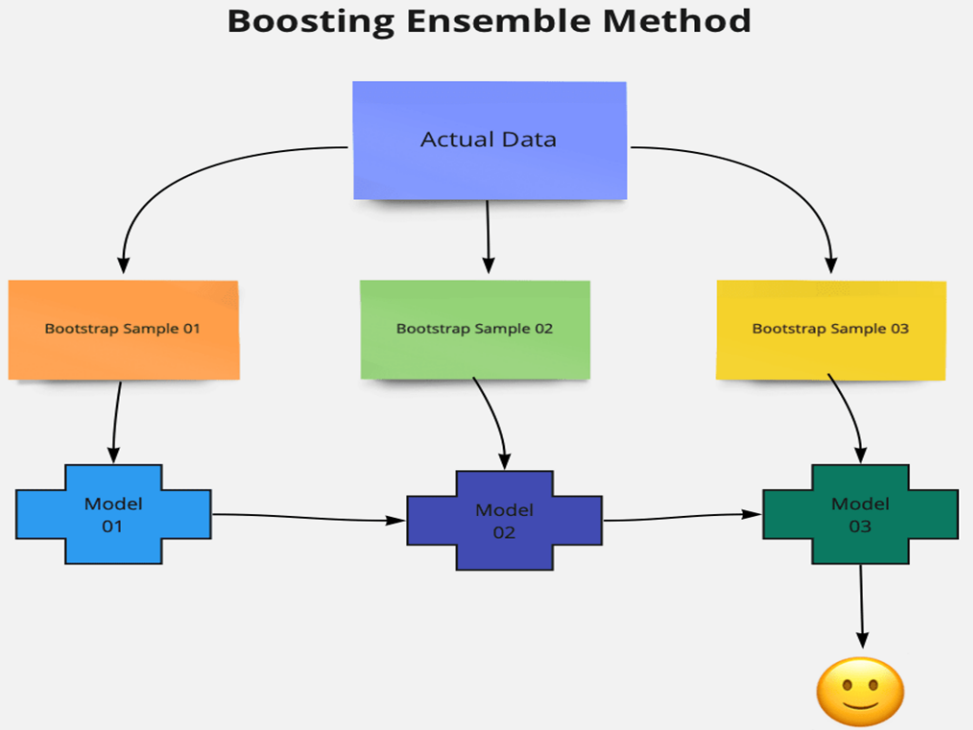


institut d'intelligence artificielle appliquée

BAGGING VS. BOOSTING



VS



dataaspirant.com

institut d'intelligence artificielle appliquée

PRINCIPAIS MÉTODOS DE ENSEMBLE

BAGGING

Processa os modelos de maneira independente paralela, e depois os combinam utilizando padrões determinísticos

(costuma considerar modelos fracos homogêneos

BOOSTING

Processa os modelos de maneira sequencial adaptativa, e depois os combinam utilizando padrões determinísticos

(costuma considerar modelos fracos homogêneos

STACKING

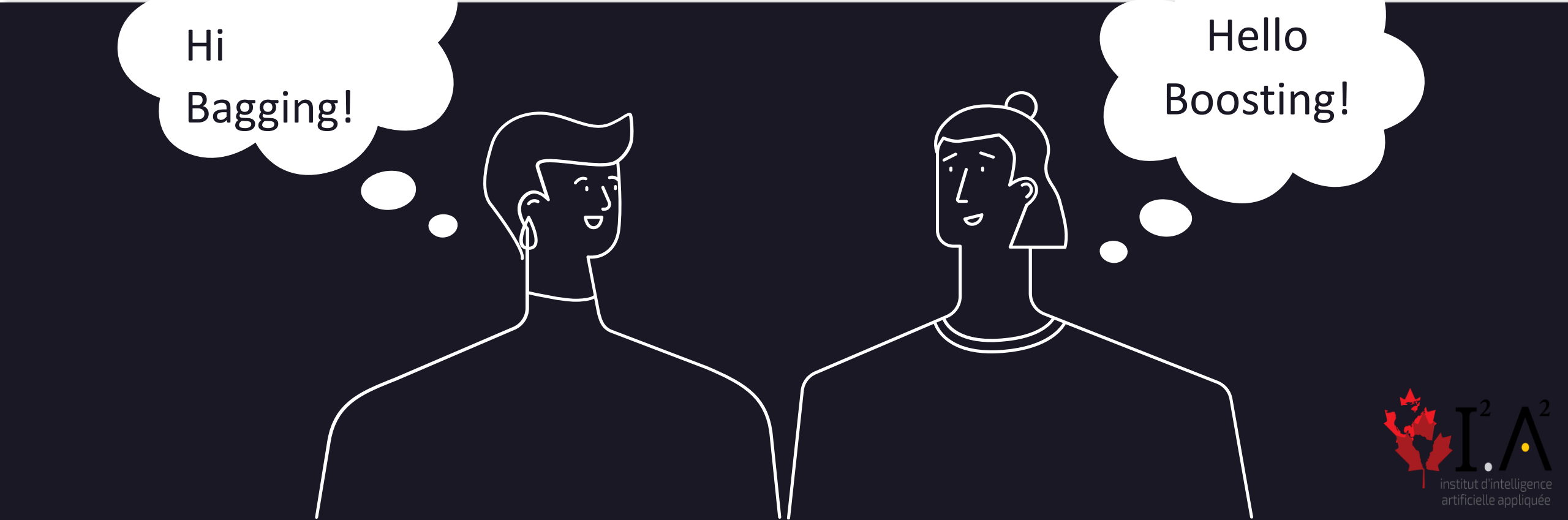
Processa os modelos de maneira paralela, e depois os combinam treinando um meta-modelo para realizar uma predição baseada em diferentes modelos fracos.

(costuma considerar modelos fracos heterogêneos

QUAL É O MELHOR?

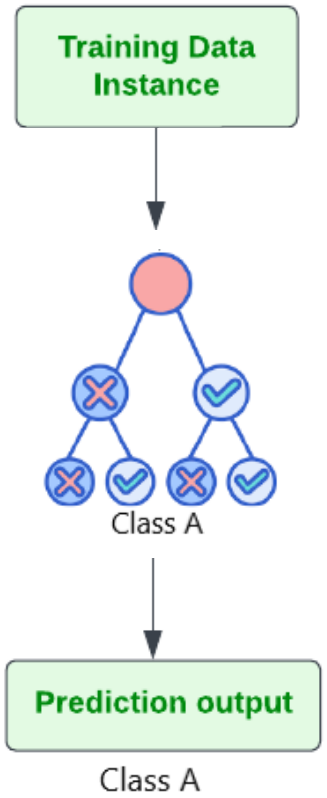
Não há melhor nem pior; depende dos dados, das simulações e das circunstâncias.

Ambos são utilizados para o mesmo princípio. Portanto, talvez o melhor seja aquele que apresente a melhor relação entre variância e viés.



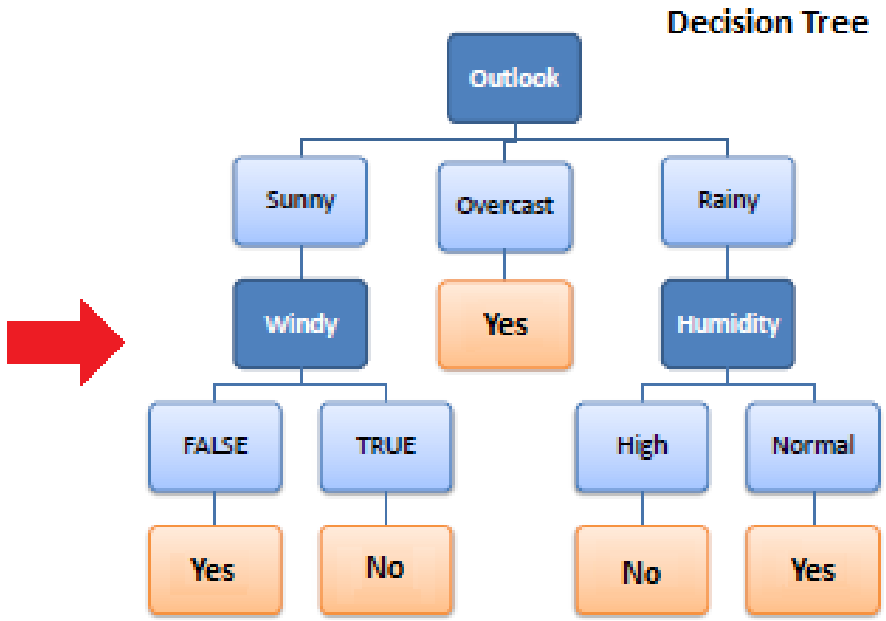
institut d'intelligence artificielle appliquée

DECISION TREE



Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



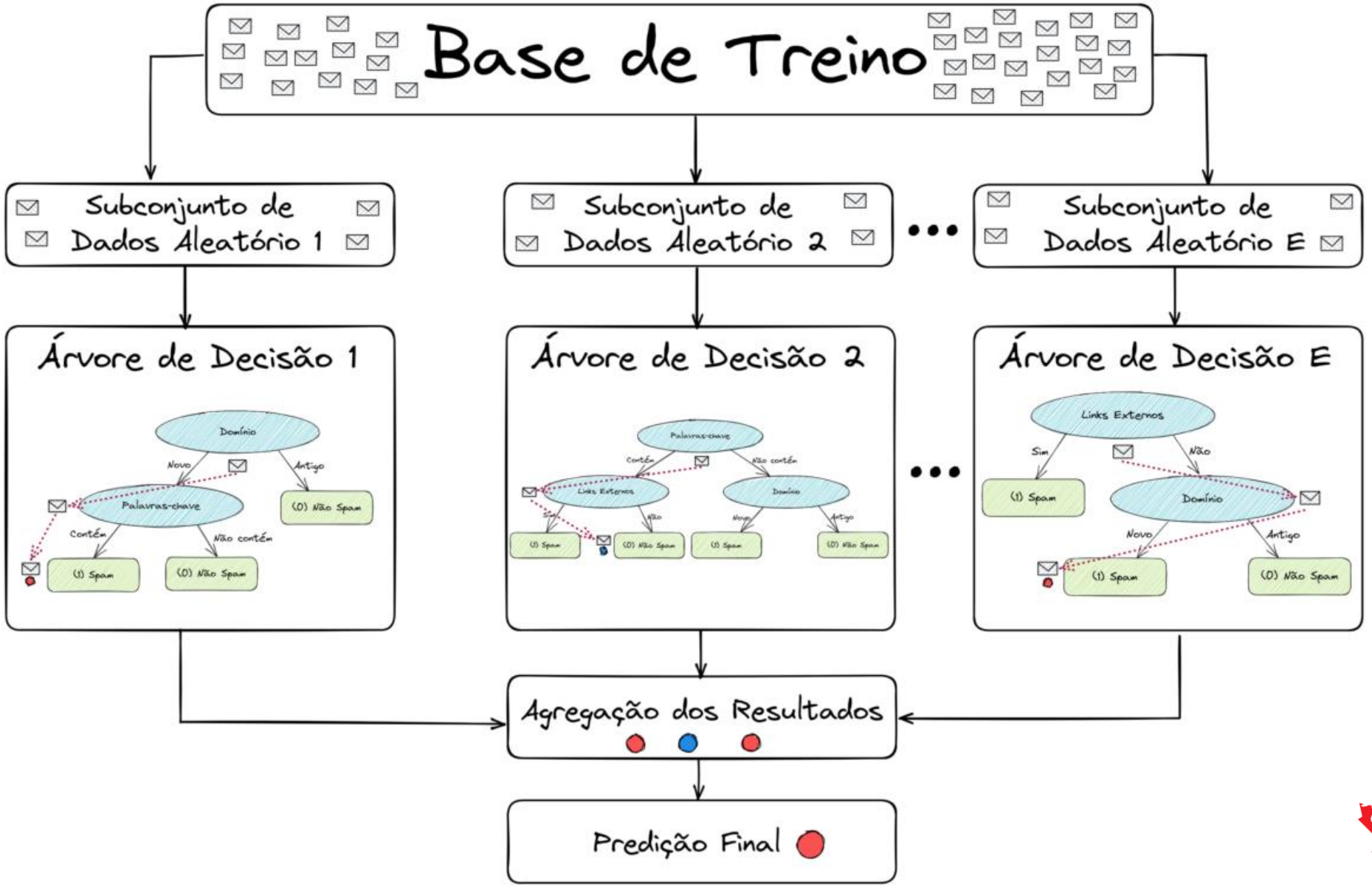
RANDOM FOREST

Extensão da Árvore de Decisão (Leo Breiman e Adele Cutler, 1996)

- Classificação e regressão
- Melhor split - Entropia, Log-loss (Ganho de informação de Shannon) e Gini

Exército de árvores de decisão

- Tomada de decisão colaborativa - "Opinião" de vários estimadores
- Treinadas com amostras aleatórias
- Considerado subconjunto de features para split - aumento da variabilidade



Fonte: https://brains.dev/2023/random-forests-algoritmos-baseados-em-arvores/

institut d'intelligence artificielle appliquée

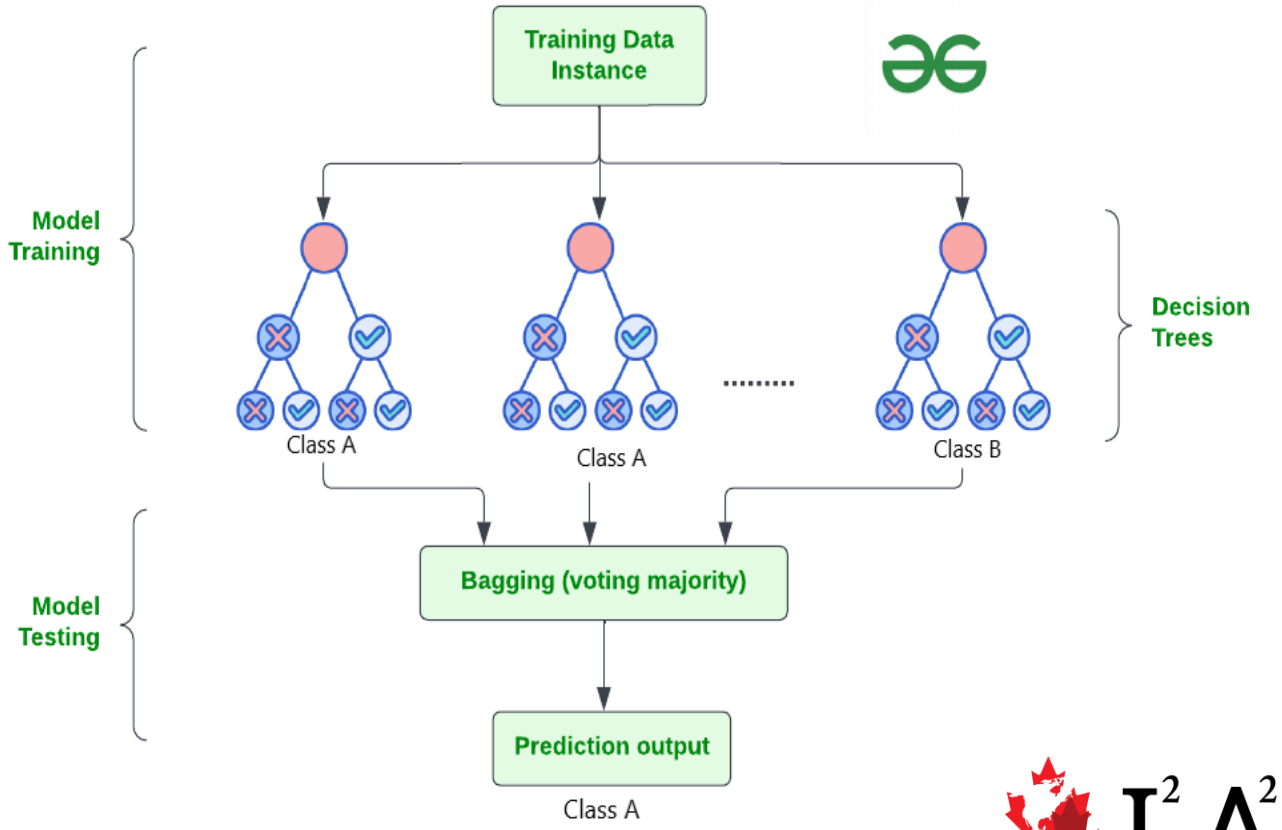
RANDOM FOREST

Pontos positivos:

- Reduz overfitting
- Performa bem com dados de alta dimensionalidade
- Paralelização
- Regressão ou classificação
- Lida bem com dados desbalanceados
- Checagem de Feature importance
- Consegue lidar com valores faltantes

Pontos negativos:

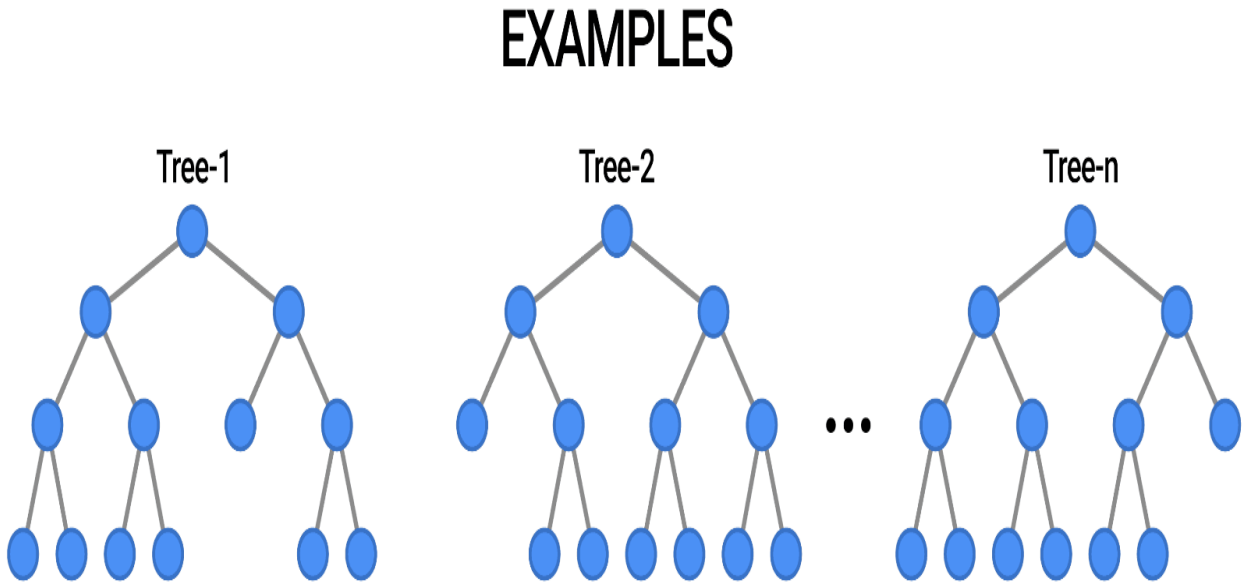
- Menor interpretabilidade
- Pode ser computacionalmente caro



institut d'intelligence artificielle appliquée

EXTRA TREES (EXTREME RANDOMIZED TREES)

- Variação de random forest proposta por Pierre Geurts, Damien Ernst e Louis Wehenkel
- Camada adicional de aleatoriedade por isso o nome "extreme":
  - Random Forest = Best Split (Entropia)
  - Extra Trees = Random Split
- Mais rápida na construção comparada a Random Forest por não buscar melhor ponto de divisão - pode resultar em maior variância sem afetar significativamente a precisão
- Combinação de classificadores fracos (baixa acurácia) podem resultar em boa acurácia/performance



institut d'intelligence artificielle appliquée

EXTRA TREES (EXTREME RANDOMIZED TREES)

Pontos positivos:

- Menor tempo de treinamento (comparada a random forest, pelo split aleatório)
- Modelo menos enviesado
- Reduz overfitting

Pontos negativos:

- Menor interpretabilidade
- Pode ter performance inferior a random forest
- Variáveis não relevantes podem influenciar na má performance do modelo (pela escolha aleatória).
- Necessário atenção ao pré-processamento para selecionar variáveis mais relevantes


OBRIGADO!
institut d'intelligence
artificielle appliquée