MBA USP ESALQ

Outros Modelos de Machine Learning II – bagging e boosting

João F. Serrajordia R. de Mello

Você vai precisar de...

Preparativos

- Abrir o R
- Importar as bibliotecas
- Algo para fazer suas anotações

Agenda

- Árvores de regressão
- *Bagging – Random Forest*
- *Boosting – Gradient Boosting*
- *Grid Search CV*

Árvores de regressão

Árvores de regressão

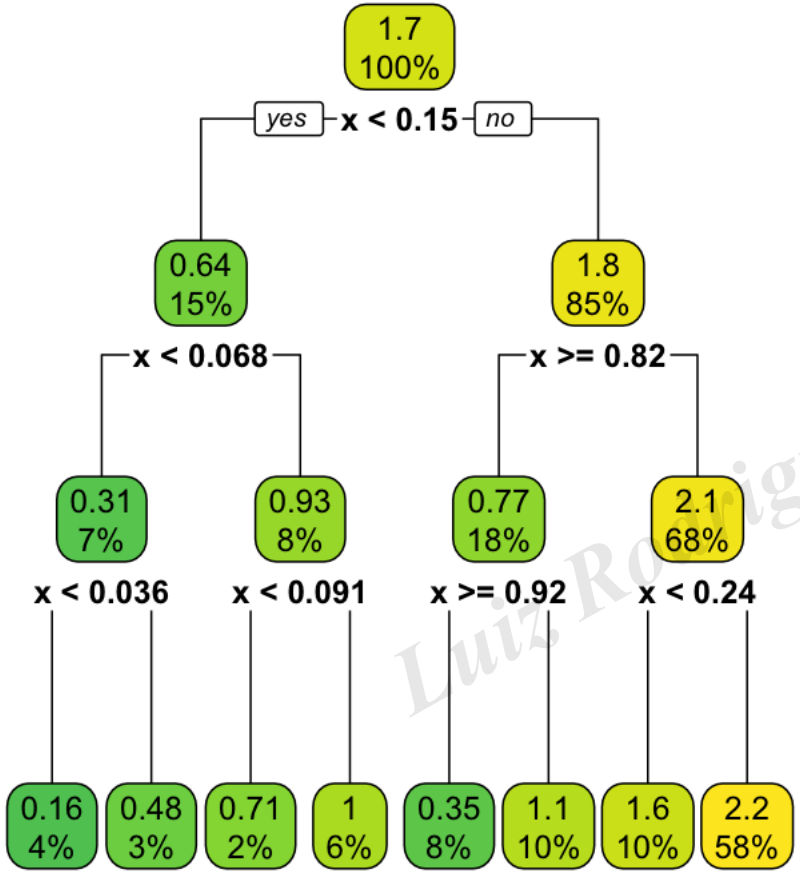
São muito semelhantes a árvores de classificação

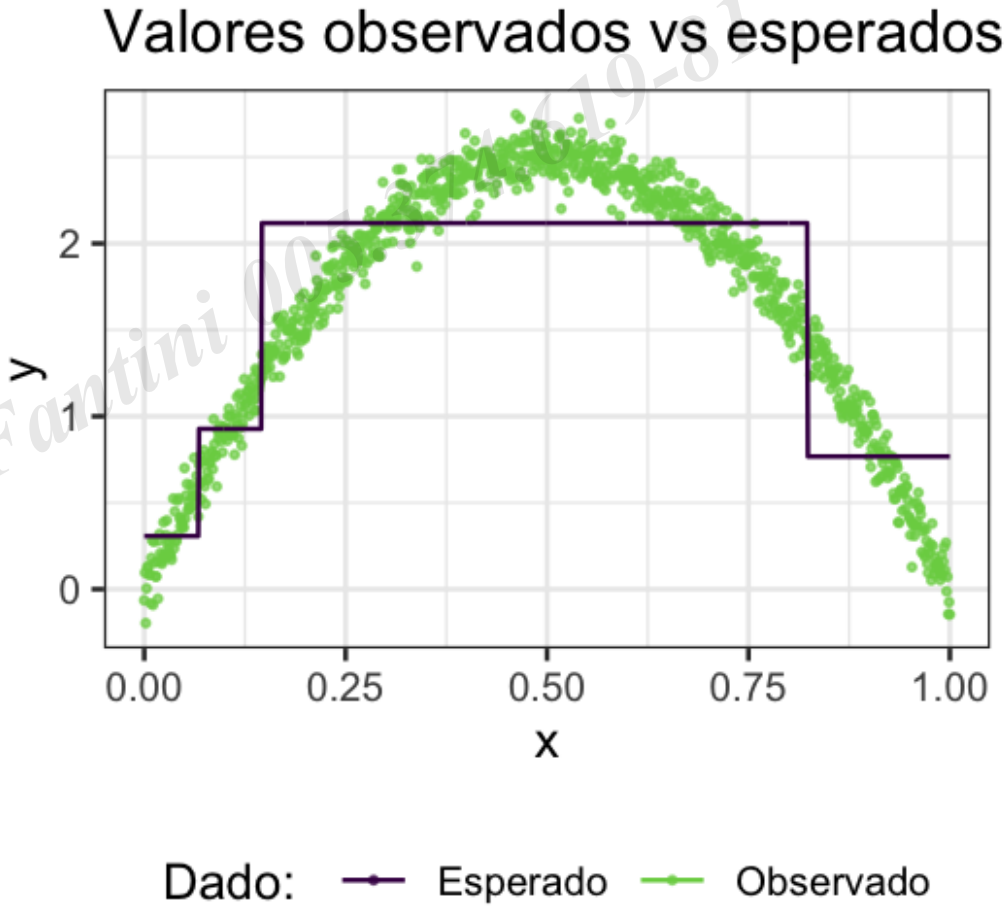
O que muda é o critério de impureza

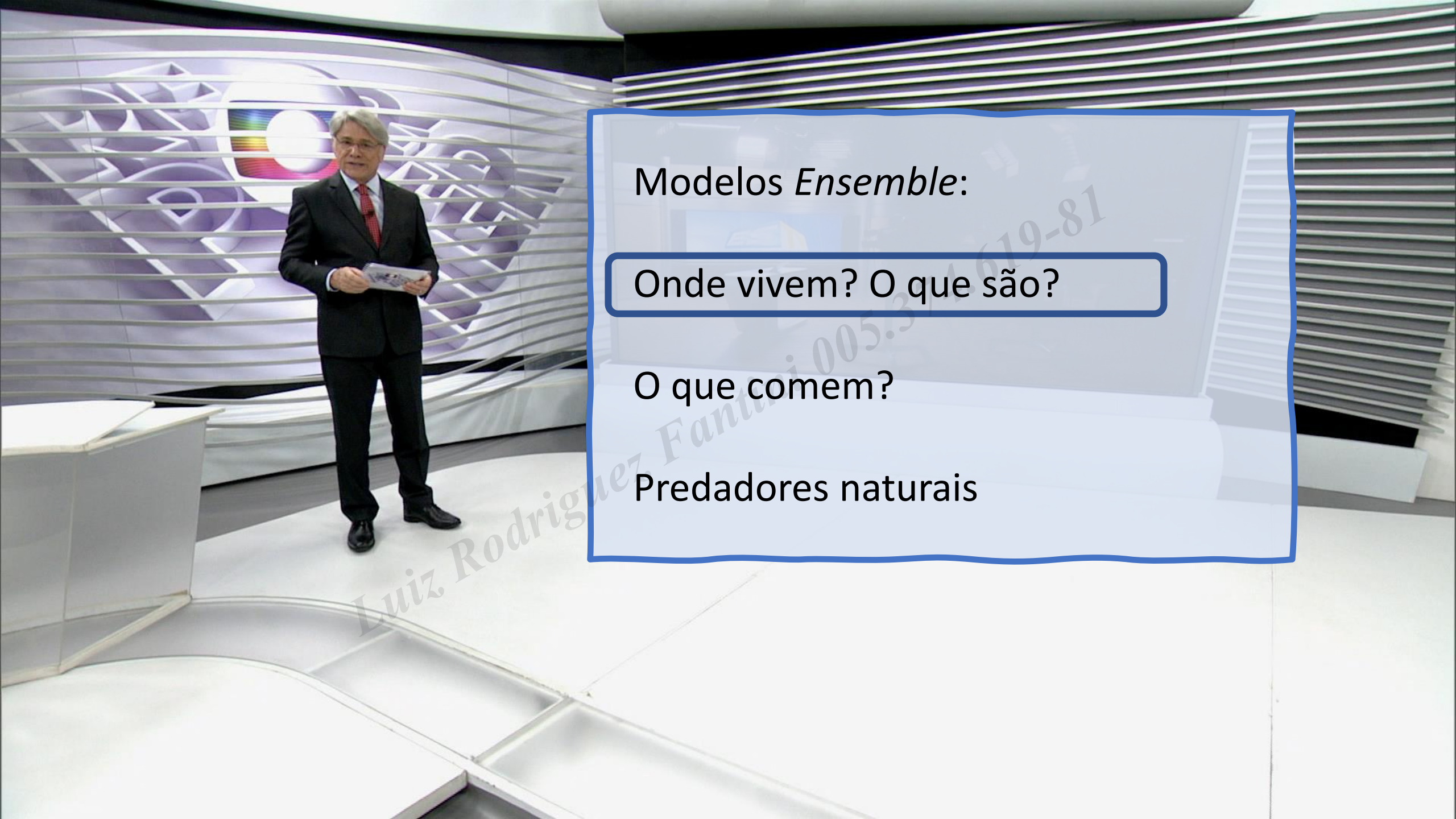
$$SQE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

https://view.genial.ly/62094a9c11415700197ff277

Árvores de regressão







Modelos *Ensemble*:

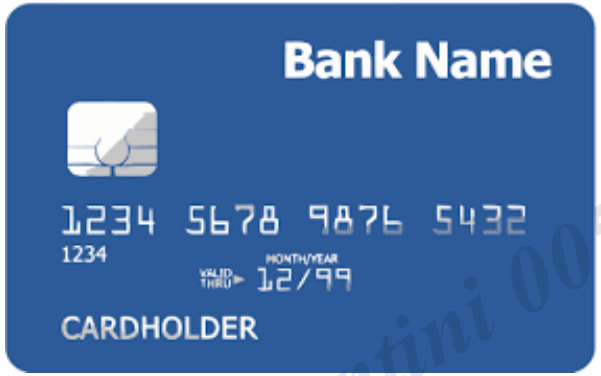
Onde vivem? O que são?

O que comem?

Predadores naturais

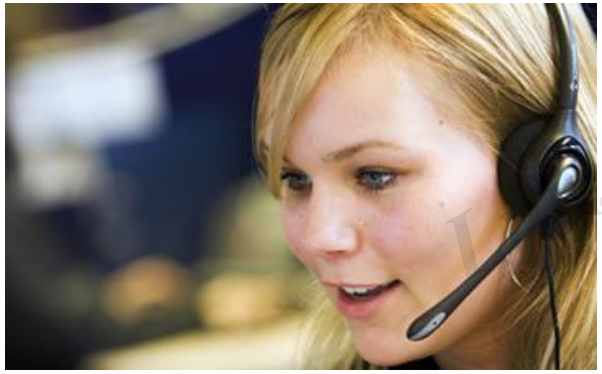
Problemas de preditivos e de classificação

Qual a eficácia de uma vacina?



O cliente vai pagar o empréstimo?

Quanto de petróleo tem no poço?

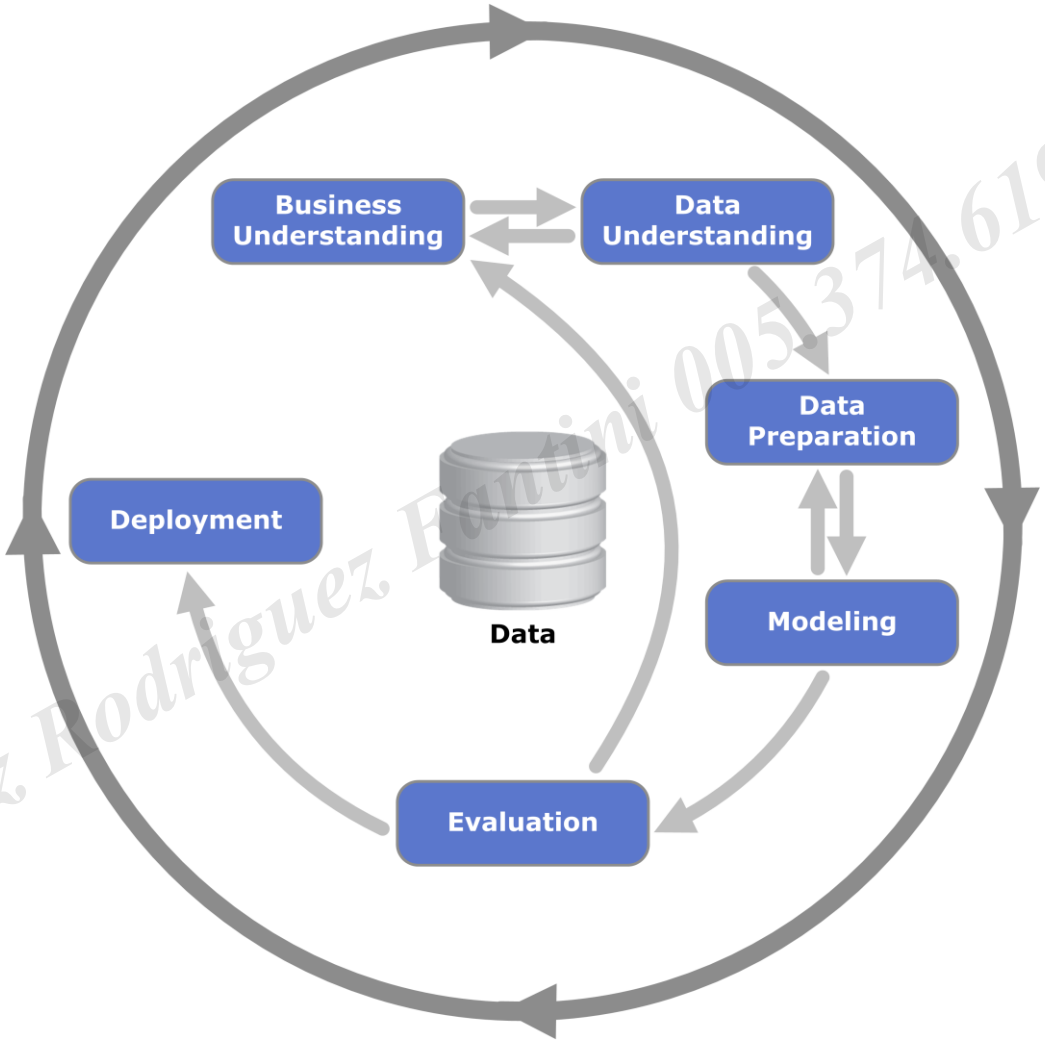


O cliente vai comprar meu produto?

O que a pessoa está fazendo?

Quão ecológico esse veículo é?

CRISP-DM



Fonte: https://www.the-modeling-agency.com/crisp-dm.pdf

Classificação dos algoritmos

Supervisionados

- Regressão
- GLM
- GLMM
- Support vector machines
- Naive Bayes
- K-nearest neighbors
- Redes Neurais
- Decision Trees

Não supervisionados

- K-Means
- Métodos hierárquicos
- Mistura Gaussiana
- DBScan
- Mini-Batch-K-Means

Estamos aqui!

Classificação dos algoritmos

Resposta contínua

- Regressão
- GLM
- GLMM
- Support vector machines
- K-nearest neighbors
- Redes Neurais
- Regression Trees

Resposta discreta

- Regressão logística
- Classification trees
- Redes Neurais
- GLM
- GLMM

Estamos aqui!

Classificação dos algoritmos

Métodos Machinelârnicos

- Árvores de decisão
- Bagging
- Boosting
- K-NN
- Redes Neurais
- Support Vector Machines

Métodos Machinelârnico-estatísticos

- Regressão
- GLM
- GLMM
- ANOVA

Estamos aqui!

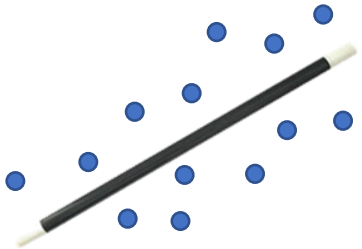
Ensemble

Um ensemble é qualquer mistura de modelos já existentes.
Os principais tipos são:

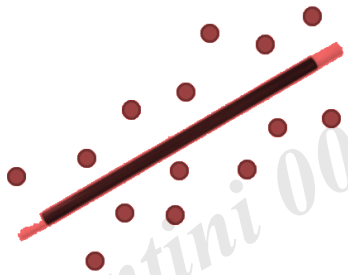
- *Bagging*
- *Boosting*
- *Stacking*

Ensemble - aggregation

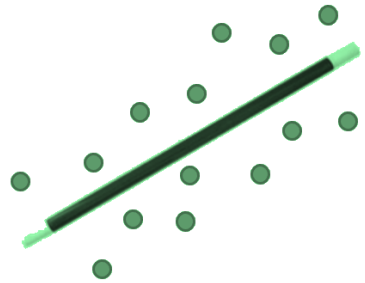
Modelo 1



Modelo 2



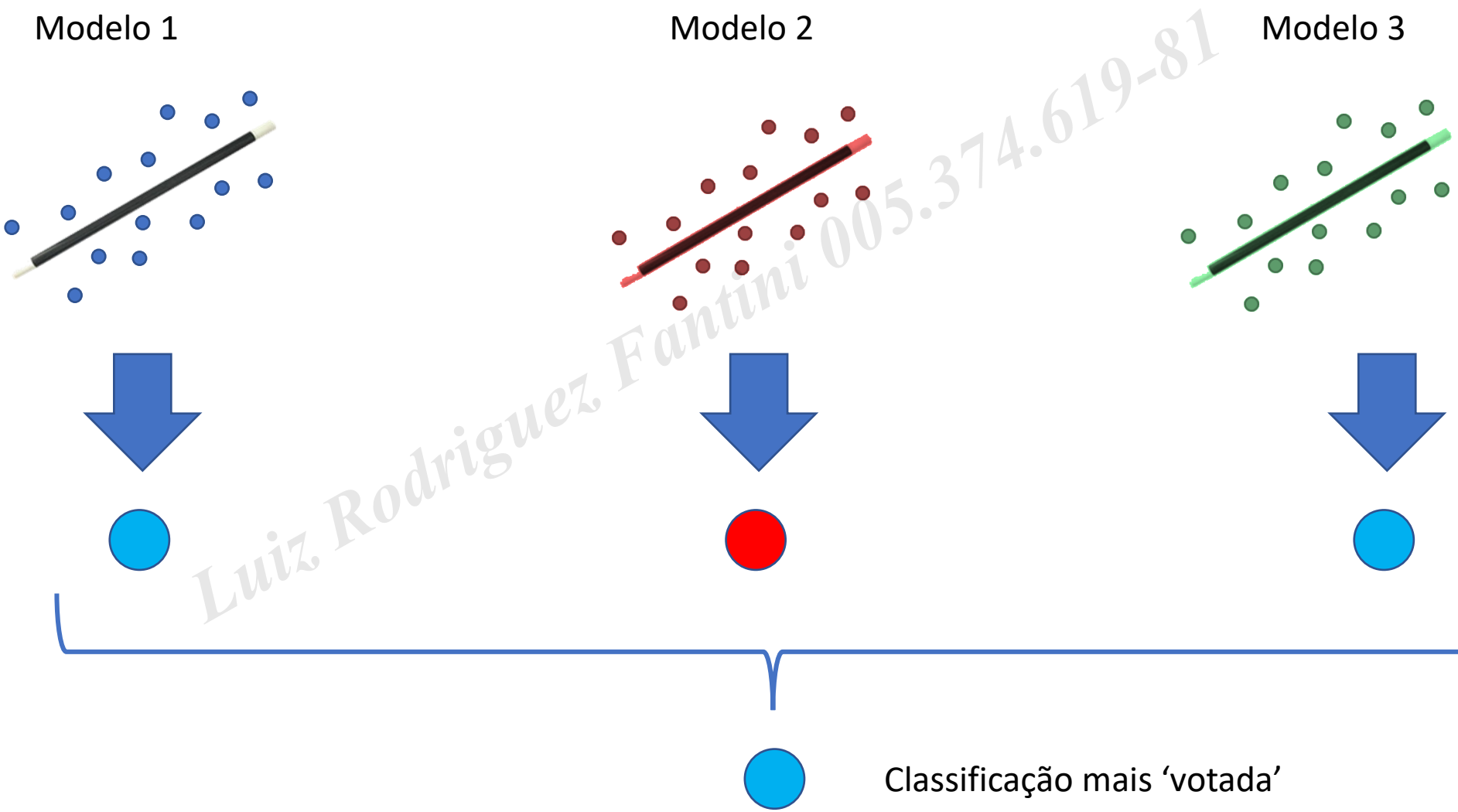
Modelo 3



Um *aggregation* consiste em uma combinação (em geral uma média simples) das previsões de dois ou mais modelos previamente construídos.

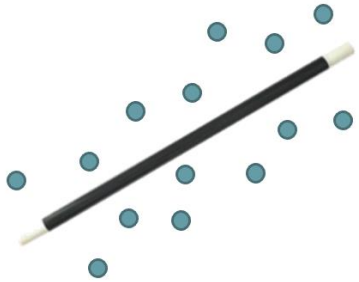
Objetivo: ainda que cada modelo seja um "*weak learner*", a combinação pode ser um "*Strong learner*" ou um preditor melhor que cada um dos integrantes.

Ensemble – Hard Voting



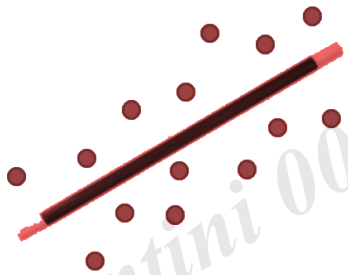
Ensemble - aggregation

Modelo 1



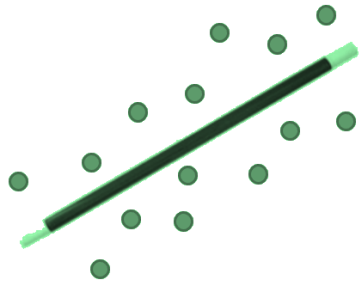
P(|) = 3%

Modelo 2



P(|) = 7%

Modelo 3



P(|) = 2%

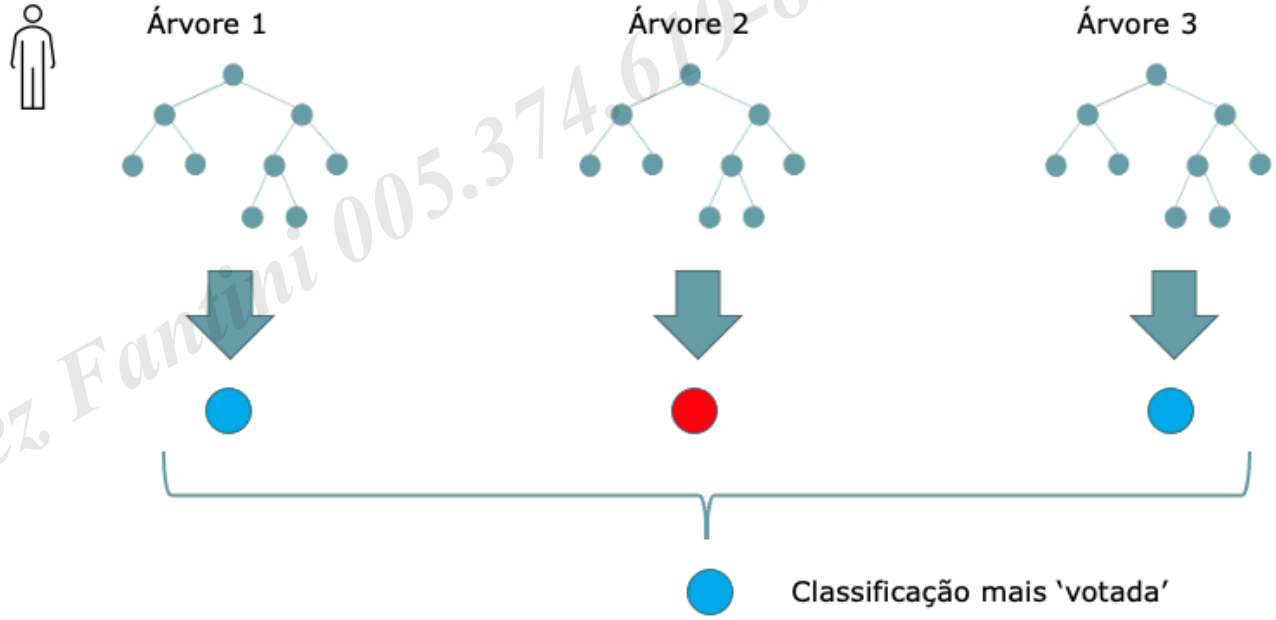
P(|) = 4%

Um método de agregação simples mas poderoso consiste em obter a média de várias previsões.

Ensemble - aggregation

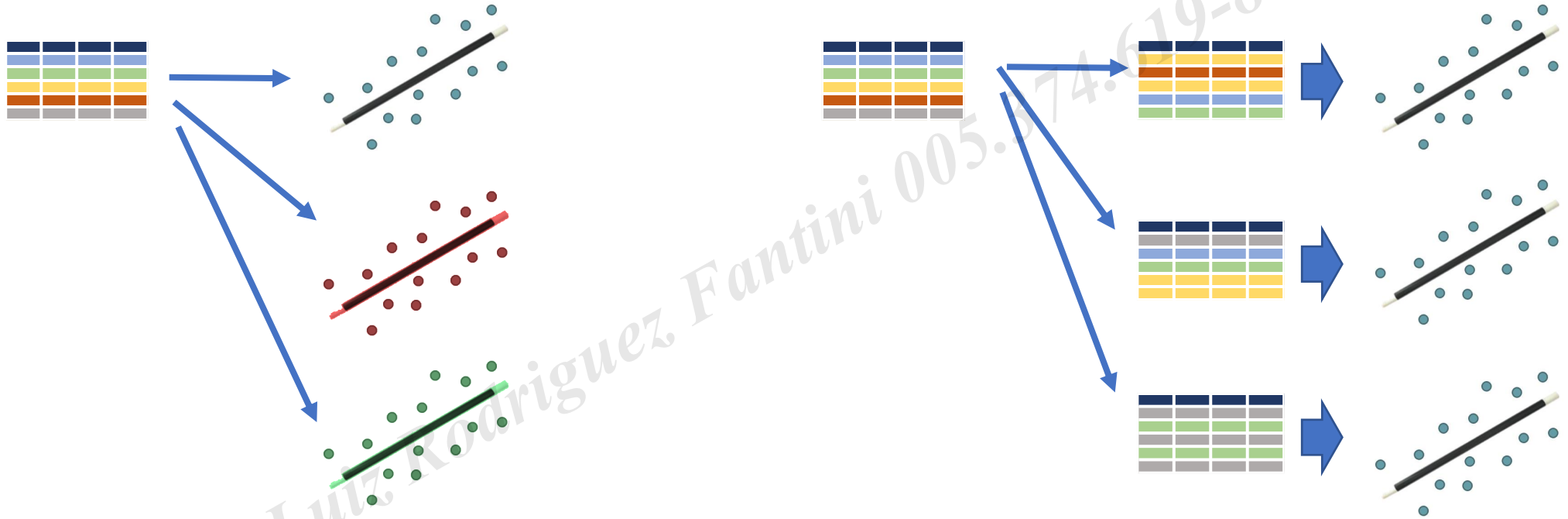
Queremos agregar modelos que sejam:

- *Úteis*
- *Mirem no mesmo alvo*
- *Diferentes*



Queremos preditores diferentes, mas que "apontem" para a mesma variável resposta. Uma ideia seria gerar preditores com alguma 'perturbação' aleatória.

Bootstrapping para avaliar a média



E se ao invés de alterar o algoritmo, alterarmos a base usando o mesmo algoritmo?

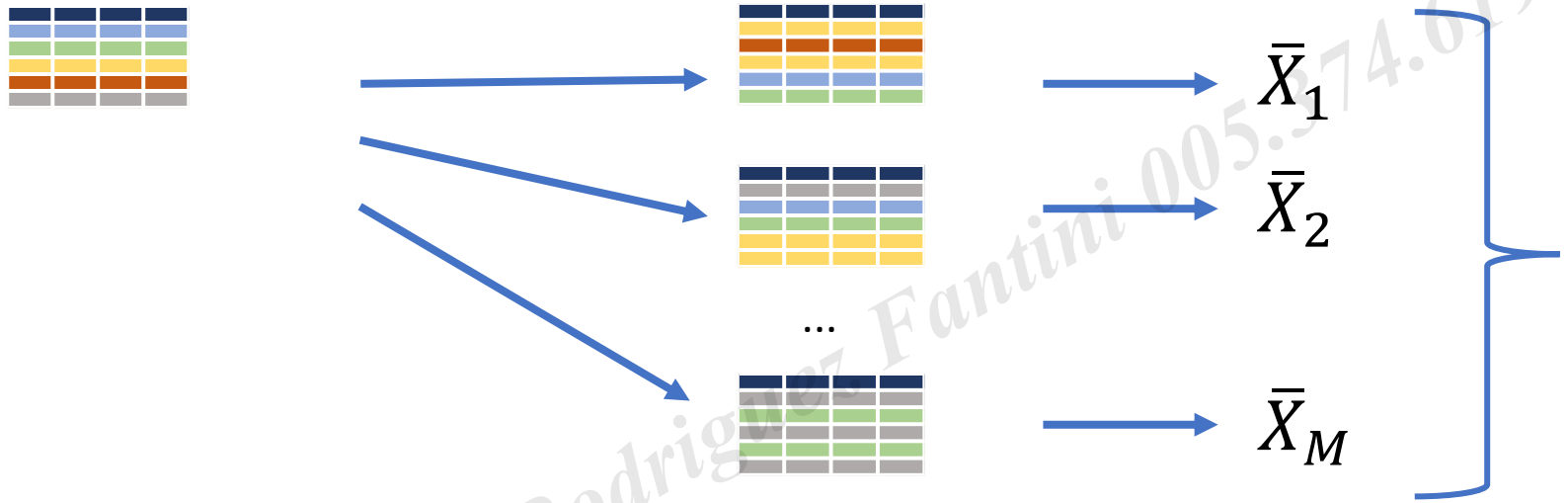
Bootstrapping para avaliar a média

Temos um conjunto de dados de tamanho N

Queremos estimar o erro padrão de um parâmetro, por exemplo, a média.

1) Retirar uma amostra aleatória de tamanho N da base
2) Calcular o parâmetro, armazenar a informação

Bootstrapping para avaliar a média

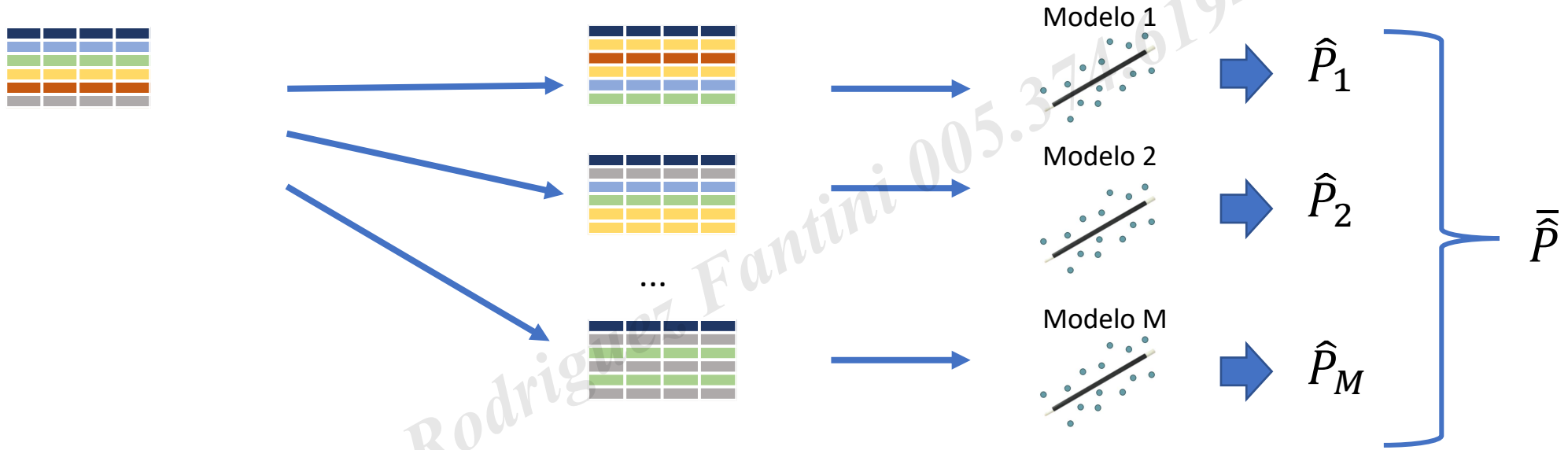


$$\bar{\bar{X}}$$

$$DesvPad(\bar{X})$$

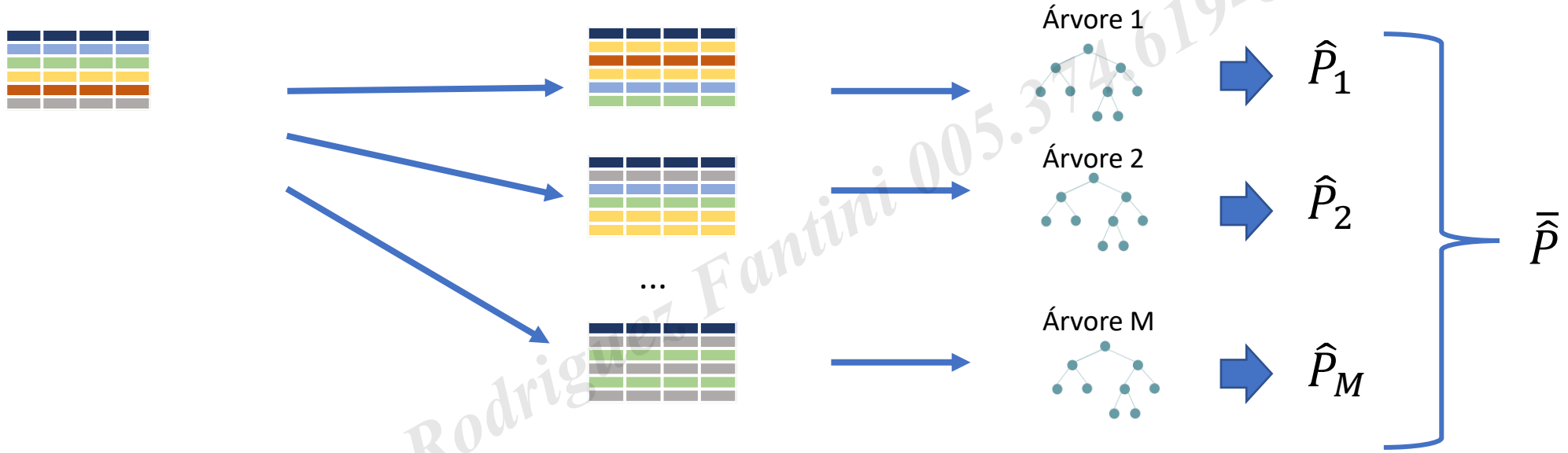
3) Repetimos isso M vezes (digamos... M=10.000 vezes)
4) Podemos calcular a média e o erro padrão do estimador

Bootstrap – aggregation (bagging)



Bagging é um *agregation* do mesmo *algoritmo* em amostras *bootstrap*

Bootstrap – aggregation (bagging)



O *bagging* com árvores é o famoso *Random Forest*


RANDOM, FORREST, RANDOM!
Random Forest

Bagging e *Pasting*

Bagging

1. Retirar uma amostra aleatória **com reposição de** tamanho N
2. Construir o modelo nessa amostra
3. Repetir 1 e 2 M vezes

Pasting

1. Retirar uma amostra aleatória **SEM reposição de** tamanho Q<N
2. Construir o modelo nessa amostra
3. Repetir 1 e 2 M vezes

O *bagging* mais famoso é *Random Forest*, que é feito com árvores, daí o nome.

Características

Bagging

1. Roda em paralelo
2. Também classifica em paralelo
3. Costuma ter bom desempenho sem grandes ajustes

Se ele fosse um carro, eu diria que é um GMC Hummer H3.

Perguntas que eu tinha quando aprendi

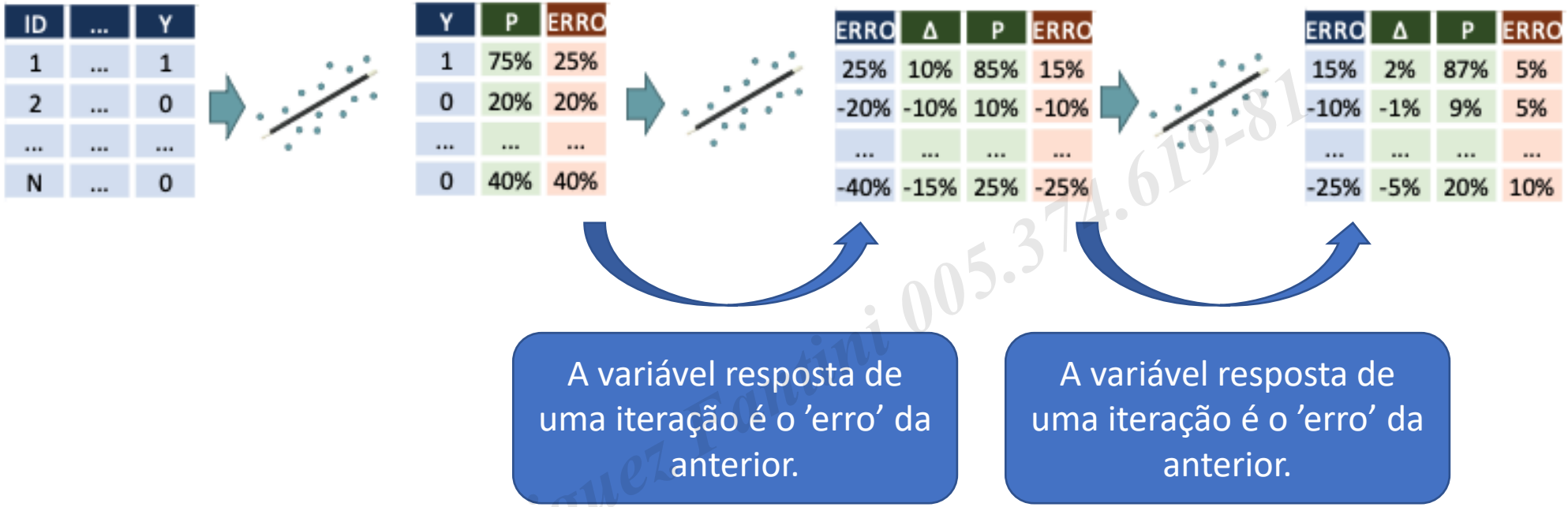
Random Forest

1. O *default* é fazer 500 árvores?
2. Demora muito para treinar?
3. E para aplicar a regra? Tenho que aplicar tudo isso de regra? Demora?
4. O algoritmo guarda tudo isso de árvore?

Se ele fosse um carro, eu diria que é um GMC Hummer H3.

Boosting

Correção sequencial de erros

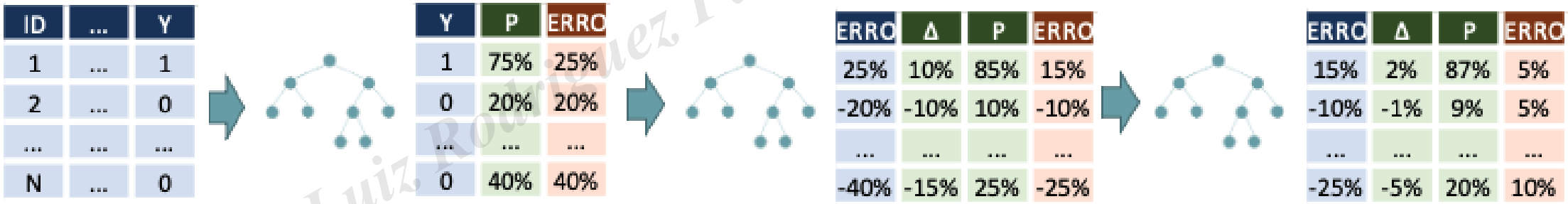


Boosting

- Os métodos de *boosting* são modelos sequenciais que tentam melhorar o erro do modelo anterior

Gradient Boosting

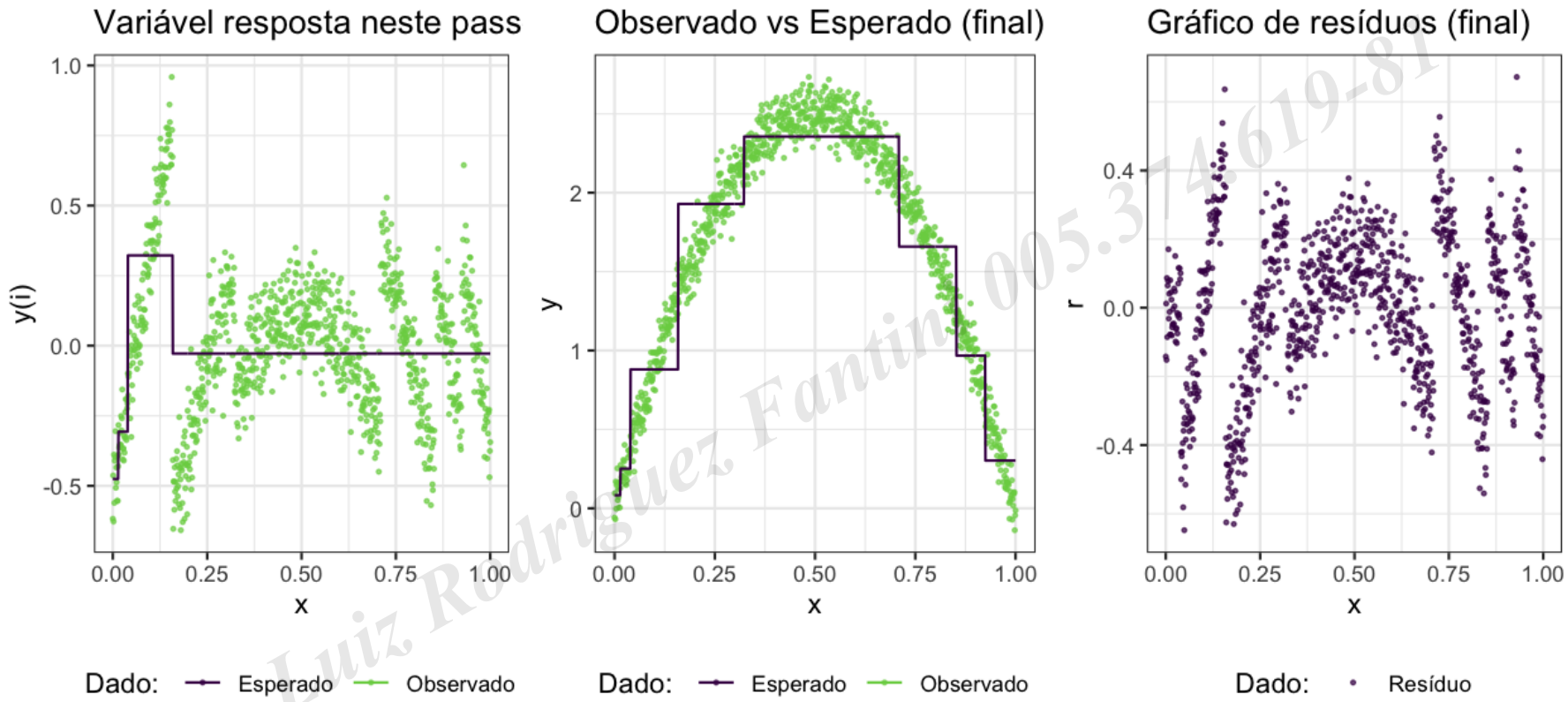
- O *Gradiente Boosting* é uma variação baseada em árvores com alguns hiperparâmetros que controlam o algoritmo



Learning rate

"Estique a corda demais e ela arrebenta, deixe-a muito frouxa, e o instrumento não toca"

Learning rate



O Learning Rate diminui o impacto de cada iteração
costuma demandar mais iterações,
mas ajuda a alcançar melhores resultados

XGBoosting

- Nome curto para Extreme Gradient Boosting
- É uma implementação do Gradient Boosting
- Possui interfaces para R e Python
- Ficou famosa por ser usada por vencedores de competições
- Criado por Tianqi Chen

O que fazer com meus novos superpoderes?

- Sugestões de prática além da aula:
  - Tentar classificar atividade humana por acelerômetro e giroscópio de celular https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones
  - Identificar doença cardíaca https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Conclusões

- Árvores são só o começo
- Há INFINITAS formas de combinar modelos, essas são as mais famosas
- Esses modelos são difíceis de se interpretar
- O *cross-validation* ‘entra no lugar’ do *stepwise*
- *PRATIQUE!*

MBA USP ESALQ

Por hoje é só ;)

linkedin.com/in/joao-serrajordia