MBA USP ESALQ

Otros Modelos de Machine Learning II – bagging y boosting

João F. Serrajordia R. de Mello

Va a necesitar de...

Preparativos

- Abrir R
- Importar las bibliotecas
- Algo para hacer sus anotaciones

Agenda

- Árboles de regresión
- *Bagging – Random Forest*
- *Boosting – Gradient Boosting*
- *Grid Search CV*

Árboles de regresión

Árboles de regresión

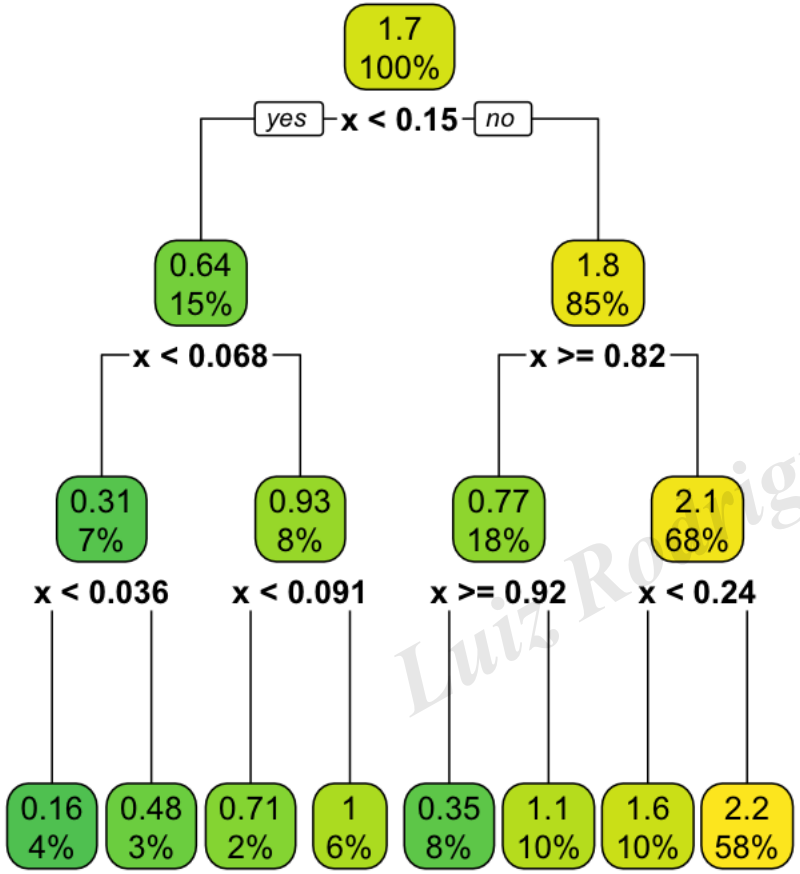
Son muy semejantes a los árboles de clasificación

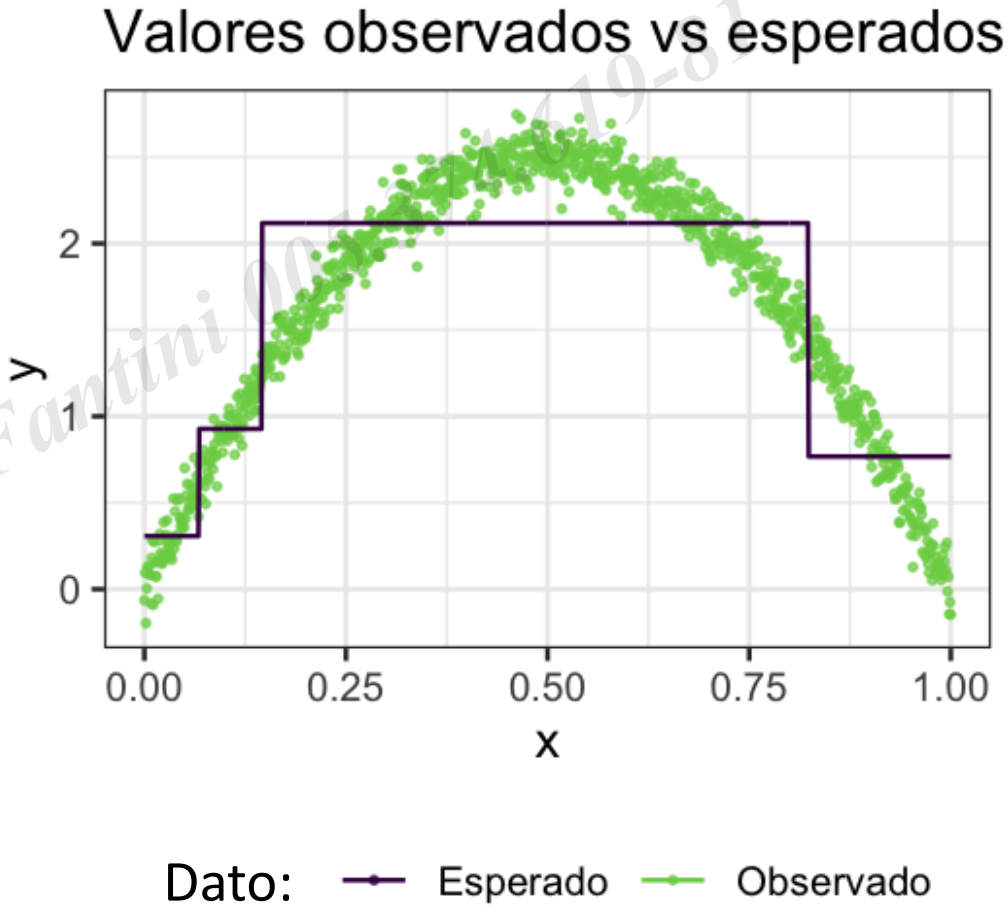
Lo que cambia es el criterio de impureza

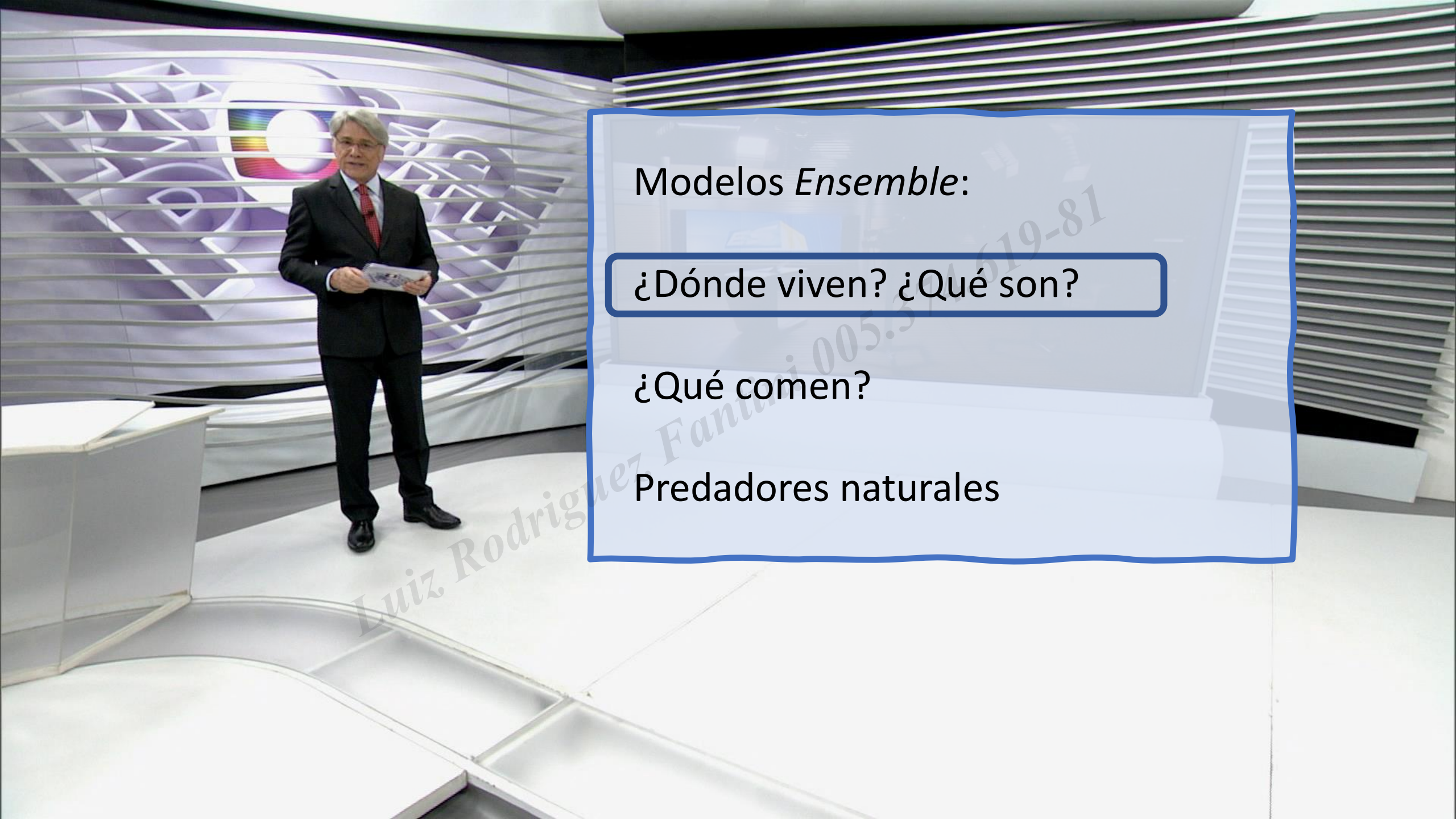
$$SQE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

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

Árboles de regresión







Modelos *Ensemble*:

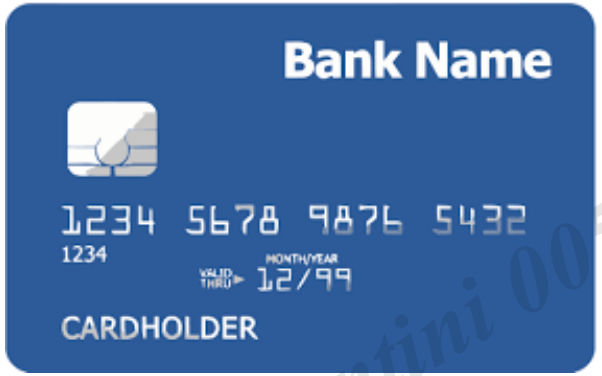
¿Dónde viven? ¿Qué son?

¿Qué comen?

Predadores naturales

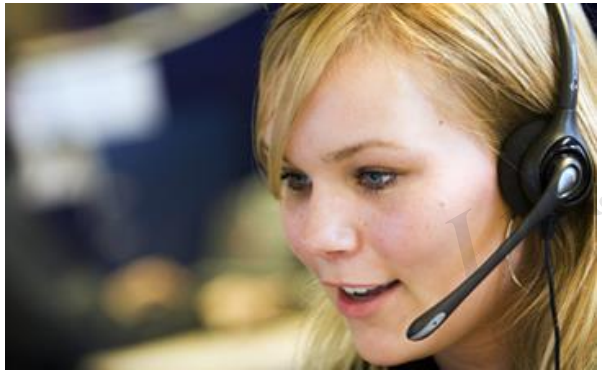
Problemas de predictivos y de clasificación

¿Cuál es la eficacia de una vacuna?



¿El cliente pagará el préstamo?

¿Cuánto petróleo tiene el pozo?

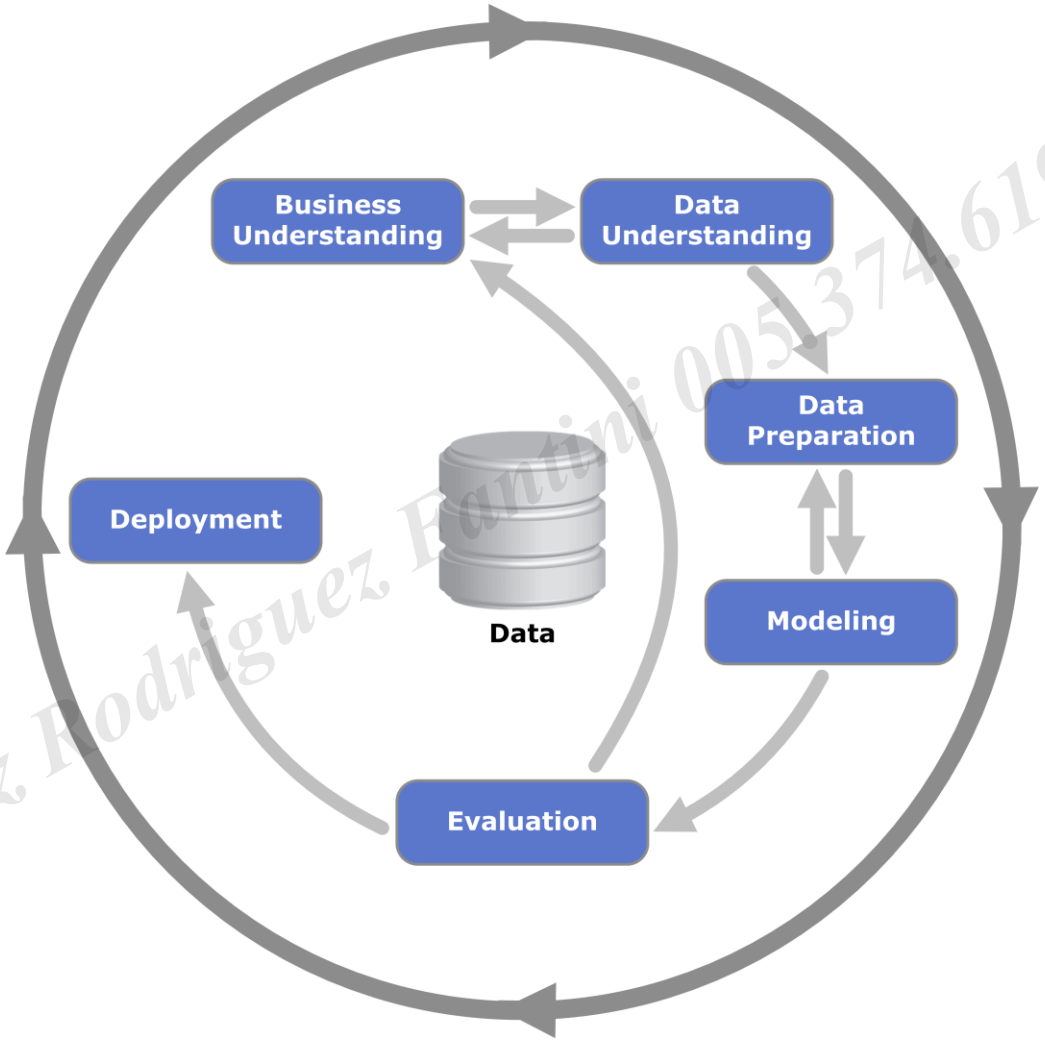


¿El cliente va a comprar mi producto?

¿Qué está haciendo la persona?

¿Cuán ecológico es ese vehículo?

CRISP-DM



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

Clasificación de los algoritmos

Supervisados

- Regresión
- GLM
- GLMM
- Support vector machines
- Naive Bayes
- K-nearest neighbors
- Redes Neurales
- Decision Trees

No Supervisados

- K-Means
- Métodos jerárquicos
- Mezcla Gaussiana
- DBScan
- Mini-Batch-K-Means

¡Estamos aquí!

Clasificación de los algoritmos

Respuesta continua

- Regresión
- GLM
- GLMM
- Support vector machines
- K-nearest neighbors
- Redes Neurales
- Regression Trees

Respuesta discreta

- Regresión logística
- Clasification trees
- Redes Neurales
- GLM
- GLMM

¡Estamos aquí!

Clasificación de los algoritmos

Métodos Machinelárnicos

- Árboles de decisión
- Bagging
- Boosting
- K-NN
- Redes Neurales
- Support vector machines

Métodos Machinelárnico-estadísticos

- Regresión
- GLM
- GLMM
- ANOVA

¡Estamos aquí!

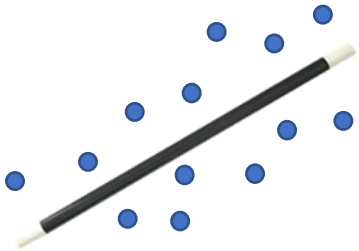
Ensemble

Un ensemble es cualquier mezcla de modelos ya existentes.
Los principales tipos son:

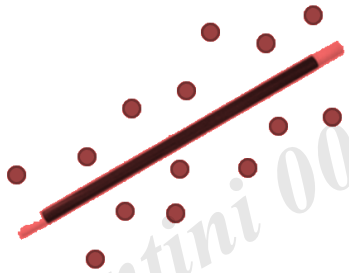
- *Bagging*
- *Boosting*
- *Stacking*

Ensemble - aggregation

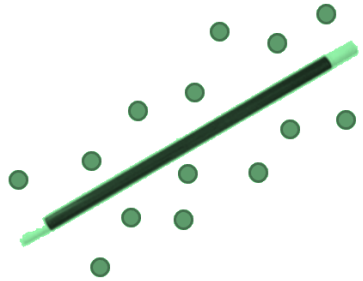
Modelo 1



Modelo 2



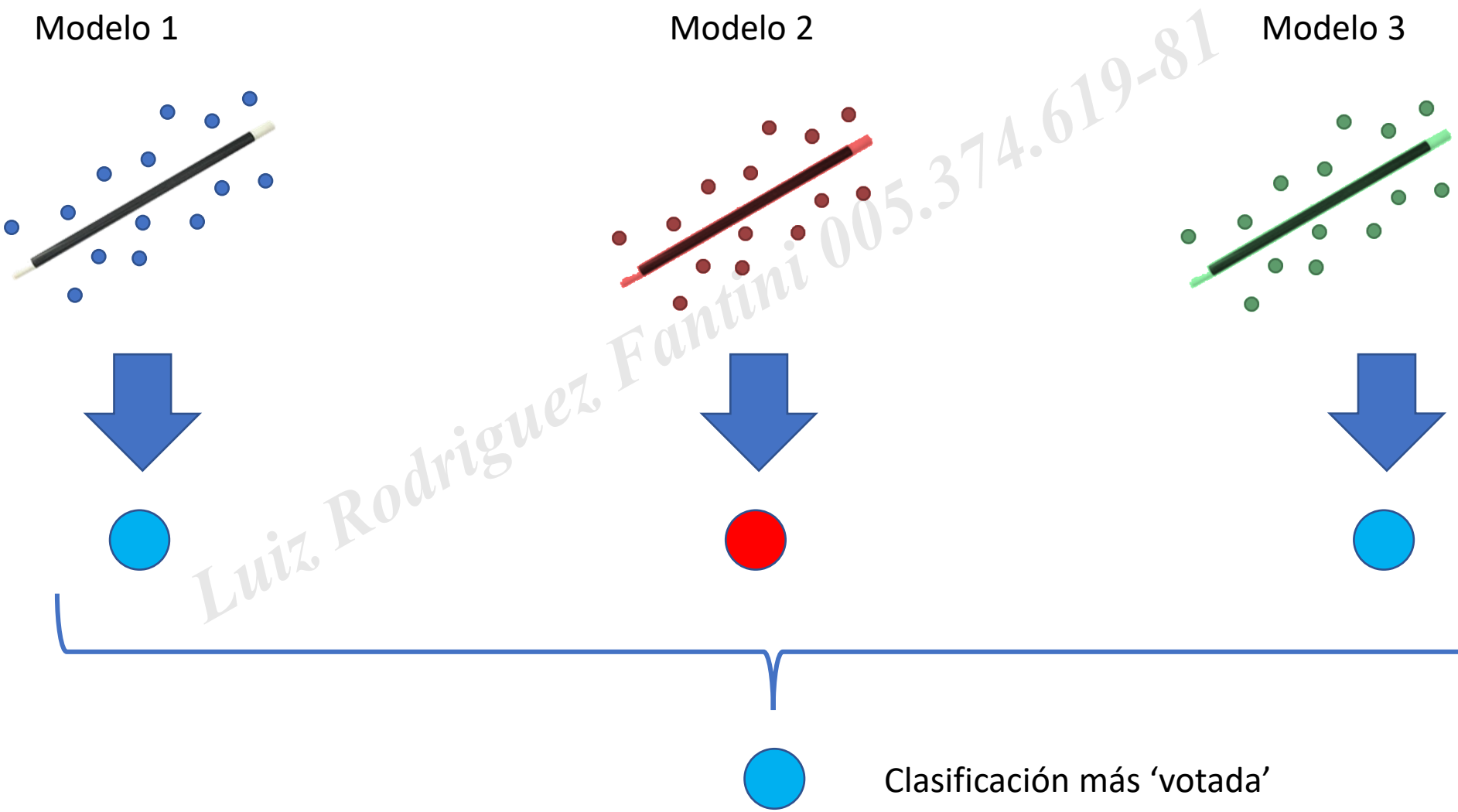
Modelo 3



Un *aggregation* consiste en una combinación (en general una media simple) de las previsiones de dos o más modelos previamente construidos.

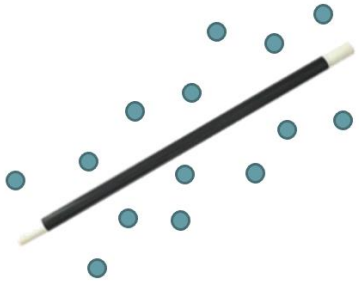
Objetivo: aun siendo cada modelo un "*weak learner*", la combinación puede ser un "*Strong learner*" o un predictivo mejor que cada uno de los integrantes.

Ensemble – Hard Voting



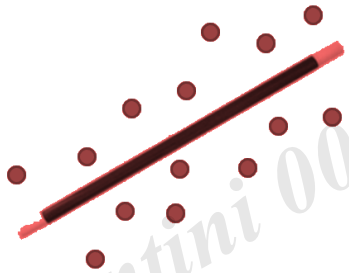
Ensemble - aggregation

Modelo 1



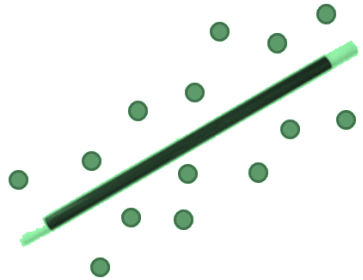
P(|) = 3%

Modelo 2



P(|) = 7%

Modelo 3



P(|) = 2%

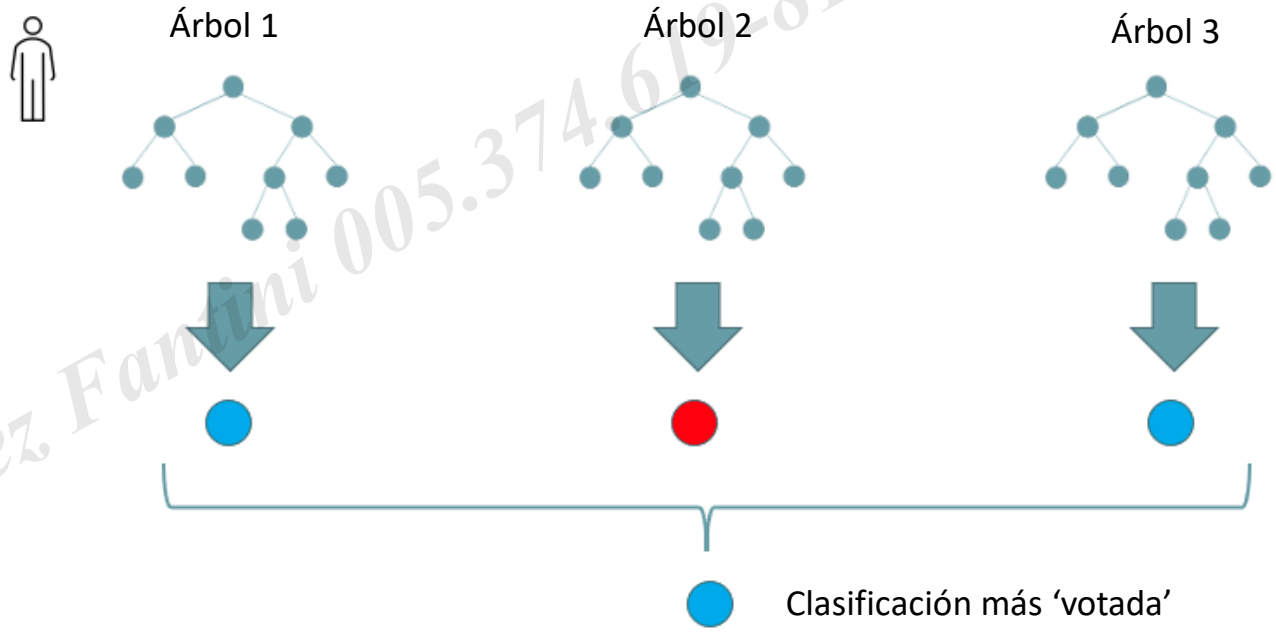
P(|) = 4%

Un método de agregación simple pero poderoso consiste en obtener la media de varias previsiones.

Ensemble - aggregation

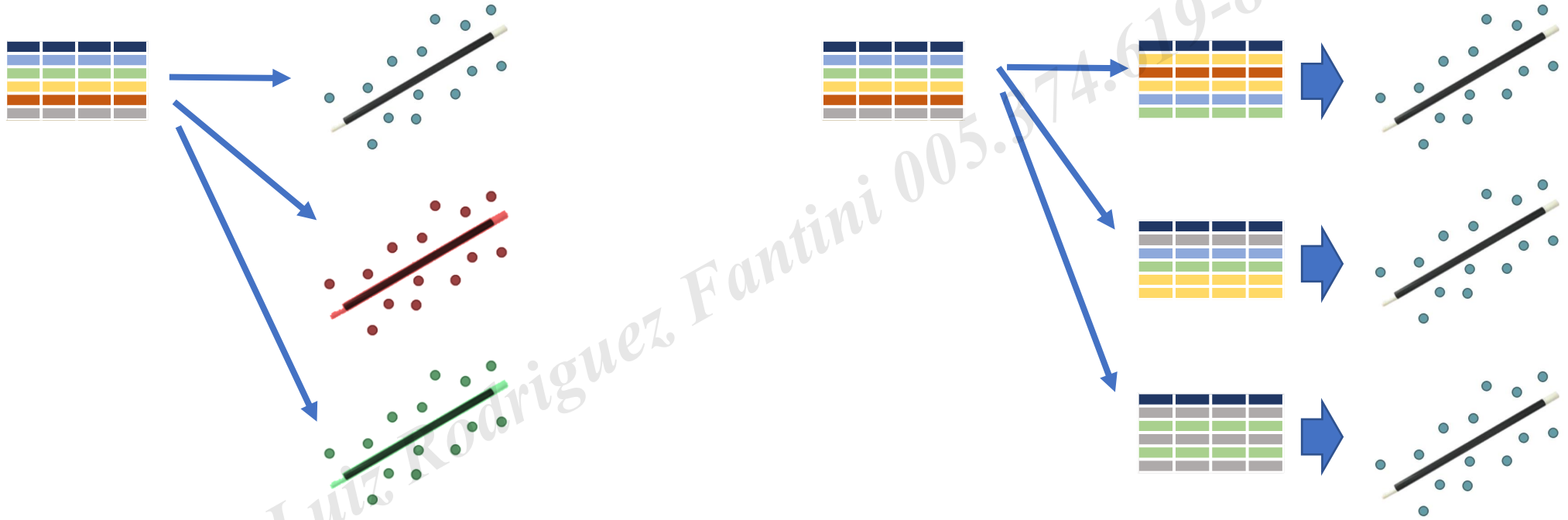
Queremos agregar modelos que sean:

- *Útiles*
- *Miren en el mismo objetivo*
- *Diferentes*



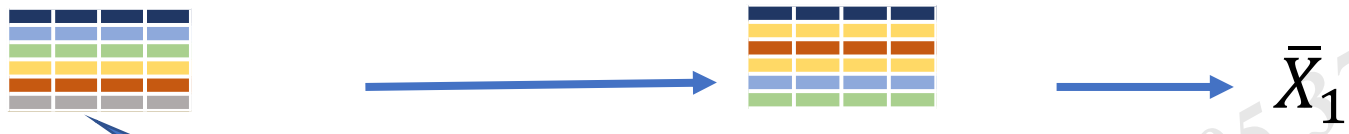
Queremos predictivos diferentes, pero que "apunten" para la misma variable respuesta. Una idea sería generar predictivos con alguna 'perturbación' aleatoria.

Bootstrapping para evaluar la media



¿Y si en vez de alterar el algoritmo, alteramos la base utilizando el mismo algoritmo?

Bootstrapping para evaluar la media

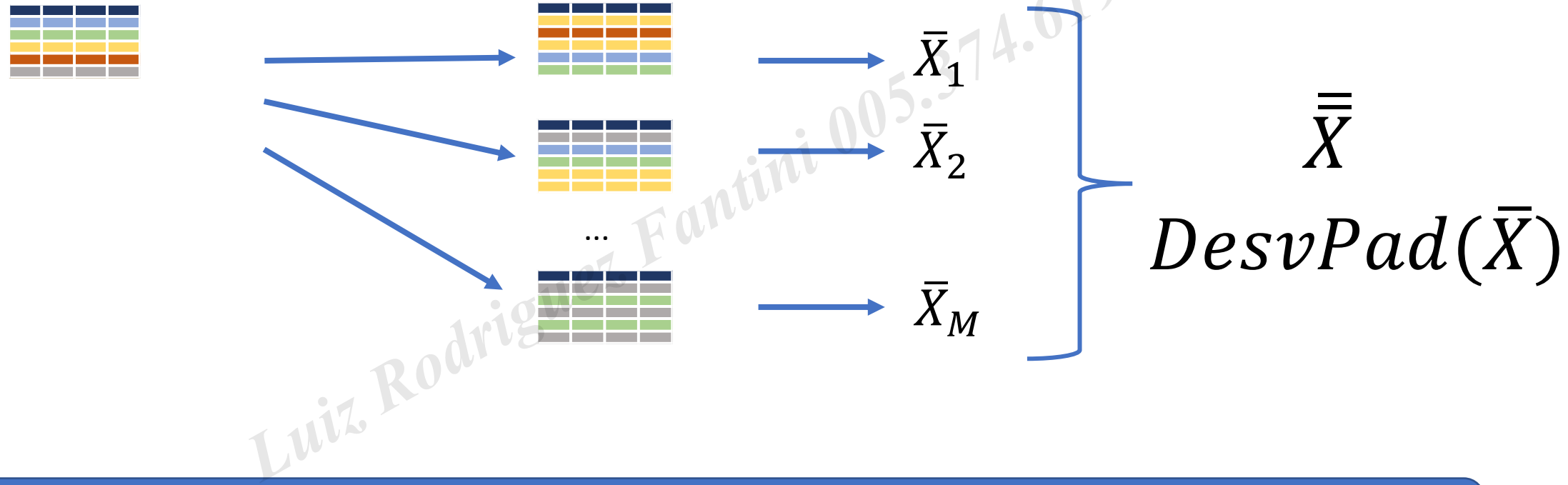


Tenemos un conjunto de datos de tamaño N

Queremos estimar el error estándar de un parámetro, por ejemplo, la media.

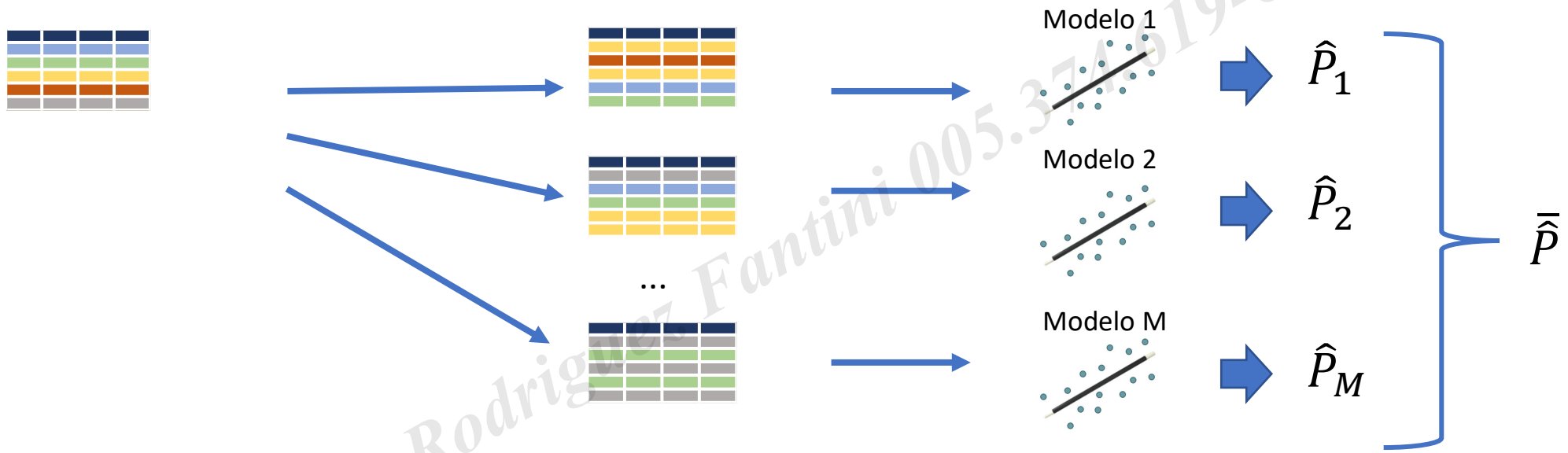
1) Retirar una muestra aleatoria de tamaño N de la base
2) Calcular el parámetro, almacenar la información

Bootstrapping para evaluar la media



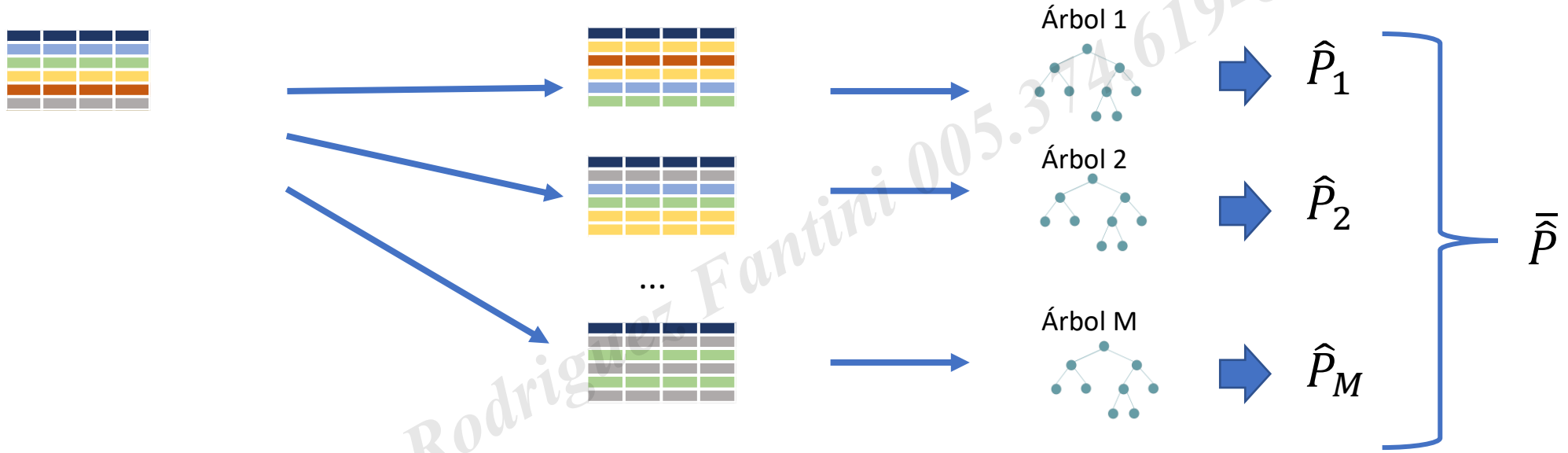
3) Repetimos eso M veces (digamos... M=10.000 veces)
4) Podemos calcular la media y el error estándar del estimador

Bootstrap – aggregation (bagging)



Bagging es un *aggregation* del mismo *algoritmo* en muestras *bootstrap*

Bootstrap – aggregation (bagging)



El *bagging* con árboles es el famoso *Random Forest*


RANDOM, FORREST, RANDOM!
Random Forest

Bagging y Pasting

Bagging

1. Retirar una muestra aleatoria **con reposición de** tamaño N
2. Construir el modelo en esa muestra
3. Repetir 1 y 2 M veces

Pasting

1. Retirar una muestra aleatoria **SIN reposición de** tamaño Q<N
2. Construir el modelo en esa muestra
3. Repetir 1 y 2 M veces

El *bagging* más famoso es *Random Forest*, que es realizado con árboles, de ahí el nombre.

Características

Bagging

1. Rueda en paralelo
2. También clasifica en paralelo
3. Normalmente tiene buen desempeño sin grandes ajustes

Si fuera un coche, diría que es un GMC Hummer H3.

Preguntas que tenía cuando aprendí

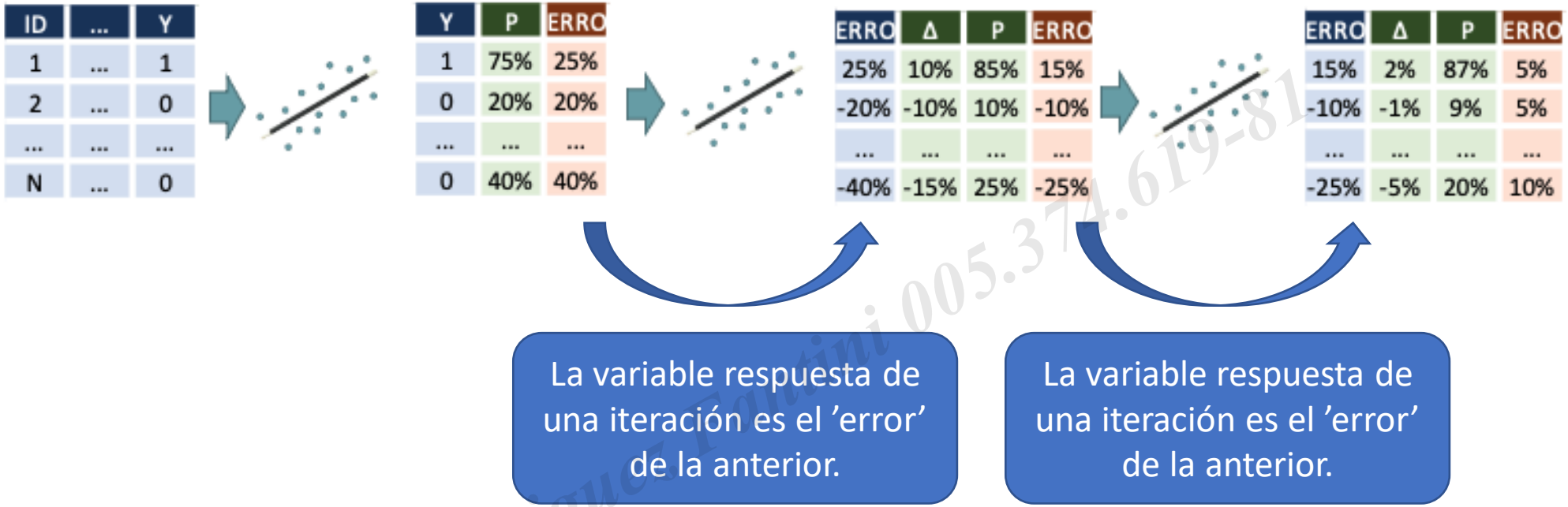
Random Forest

1. ¿*Default* es hacer 500 árboles?
2. ¿Tarda mucho para entrenar?
3. ¿Y para aplicar la regla? ¿Tengo que aplicar todo eso de regla? ¿Tarda?
4. ¿El algoritmo guarda todo eso de árbol?

Si fuera un coche, diría que es un GMC Hummer H3.

Boosting

Corrección secuencial de errores

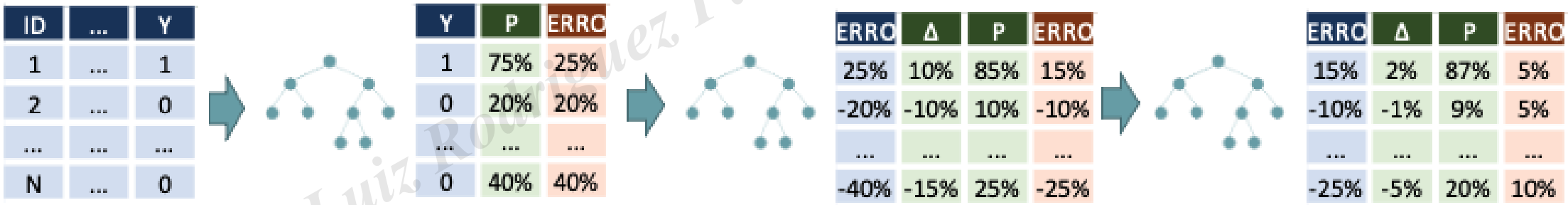


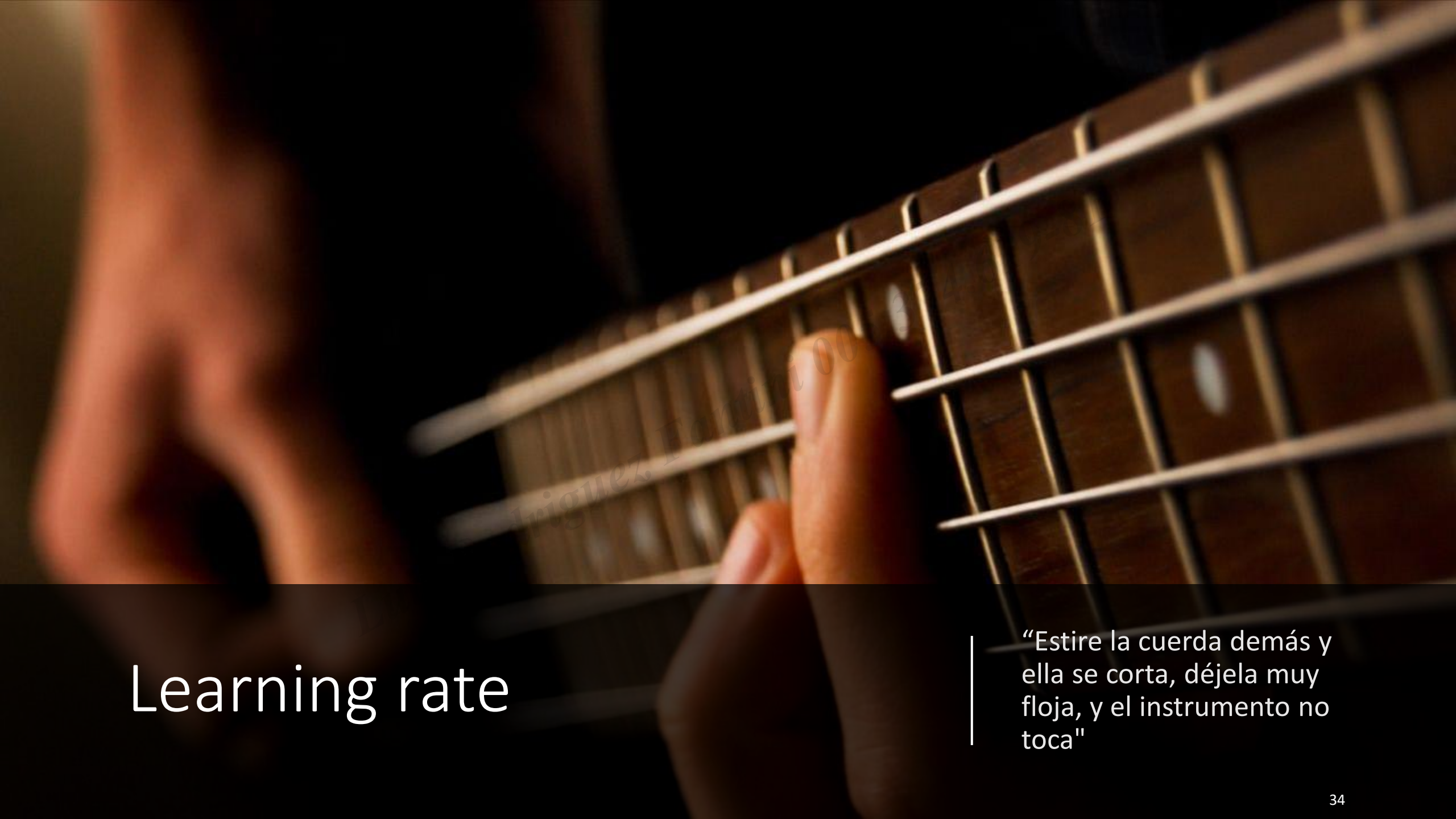
Boosting

- Los métodos de *boosting* son modelos secuenciales que intentan mejorar el error del modelo anterior

Gradient Boosting

- El *Gradiente Boosting* es una variación basada en árboles con algunos hiperparámetros que controlan el algoritmo

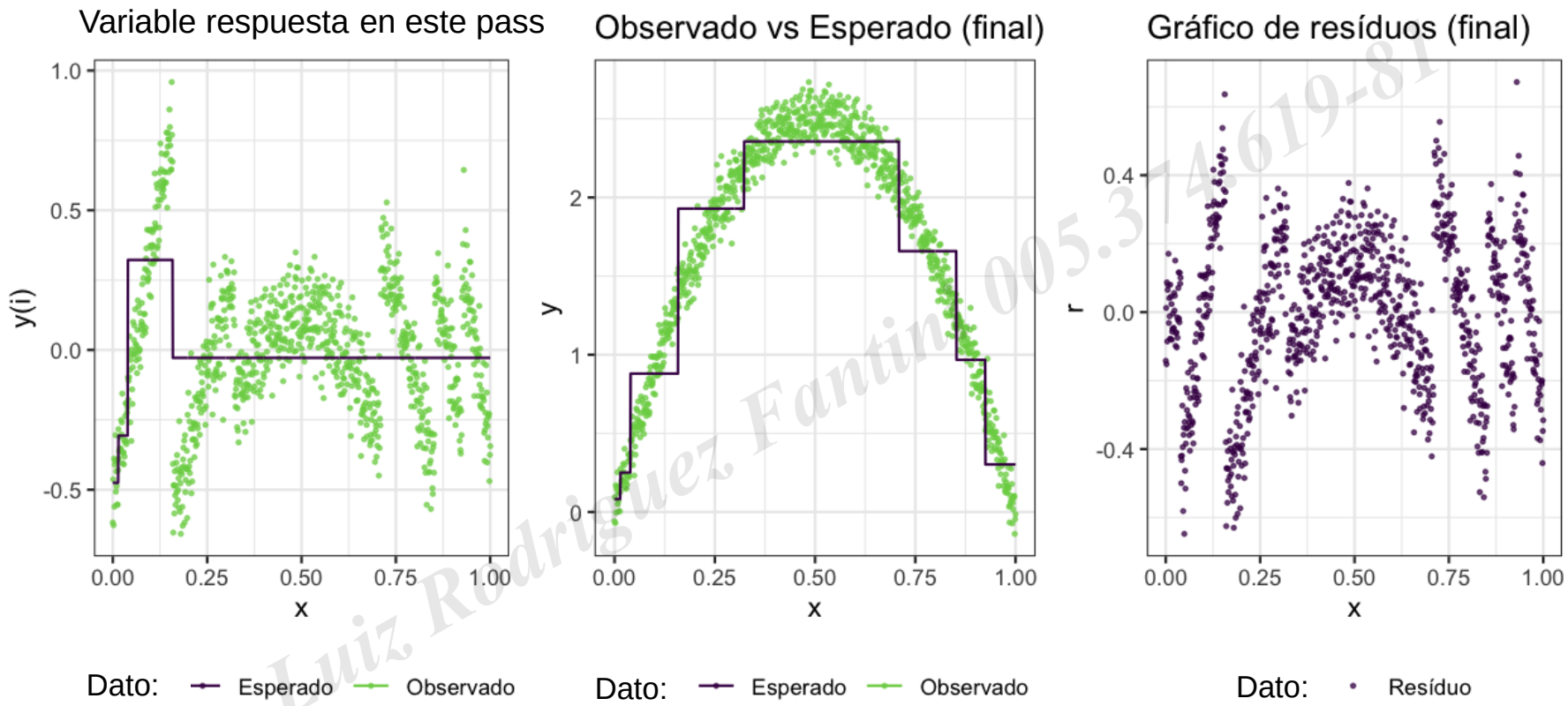




Learning rate

"Estire la cuerda demás y ella se corta, déjela muy floja, y el instrumento no toca"

Learning rate



Learning Rate disminuye el impacto de cada iteración
suele demandar más iteraciones,
pero ayuda a alcanzar mejores resultados

XGBoosting

- Nombre corto para Extreme Gradient Boosting
- Es una implementación del Gradient Boosting
- Posee interfaces para R y Python
- Se hizo famosa por ser utilizada por vencedores de competiciones
- Creado por Tianqi Chen

¿Qué hacer con mis nuevos superpoderes?

- Sugerencias de práctica además de la clase:
  - Intentar clasificar actividad humana por acelerómetro y giroscopio de celular https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones
  - Identificar enfermedad cardíaca https://archive.ics.uci.edu/ml/datas

Conclusiones

- Árboles son sólo el comienzo
- Hay INFINITAS formas de combinar modelos, esas son las más famosas
- Esos modelos son difíciles de interpretar
- El *cross-validation* ‘entra en el lugar’ del *stepwise*
- *¡PRACTIQUE!*

Por hoy es sólo eso ;)

linkedin.com/in/joao-serrajordia