De la estadística tradicional al *Machine Learning*

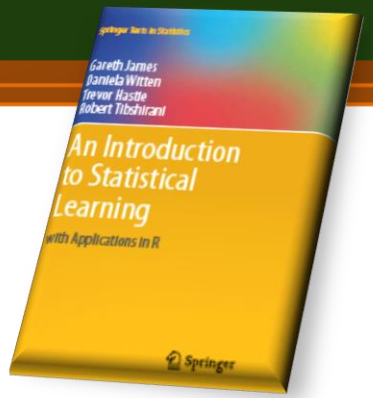
Daniel Vélez Serrano

29 mayo 2019

Un *trade-off* habitual en la modelización analítica

$Y = f(x)$

¿f?



Interpretar vs Predecir

Interpretar vs Predecir

INTERPRETAR

DEDUCIR

ENTENDER

APRENDIZAJE MANUAL

¿f?

PREDECIR

ACERTAR

¡¡¡GANAR!!!

APRENDIZAJE AUTOMÁTICO

Interpretar vs Predecir

INTERPRETAR

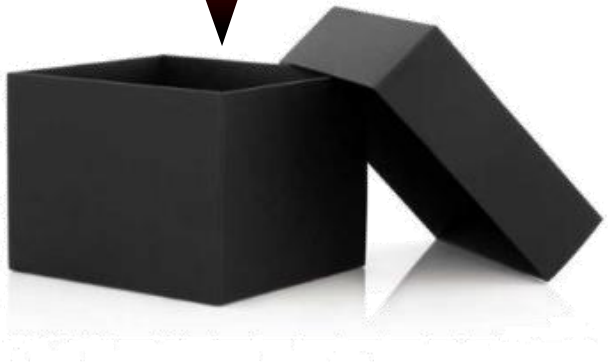
PREDECIR

interpretar Conjugar

Del lat. *interpretāri.*

1. t… Explicar … declarar el sentido de algo

¿f?

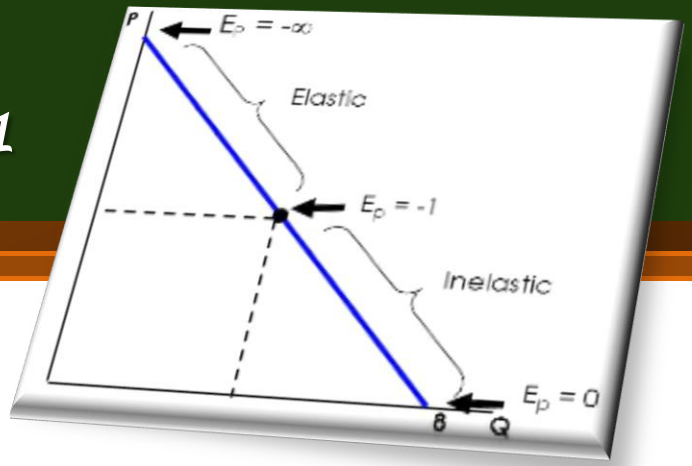


Modelización tradicional: Regresión lineal simple

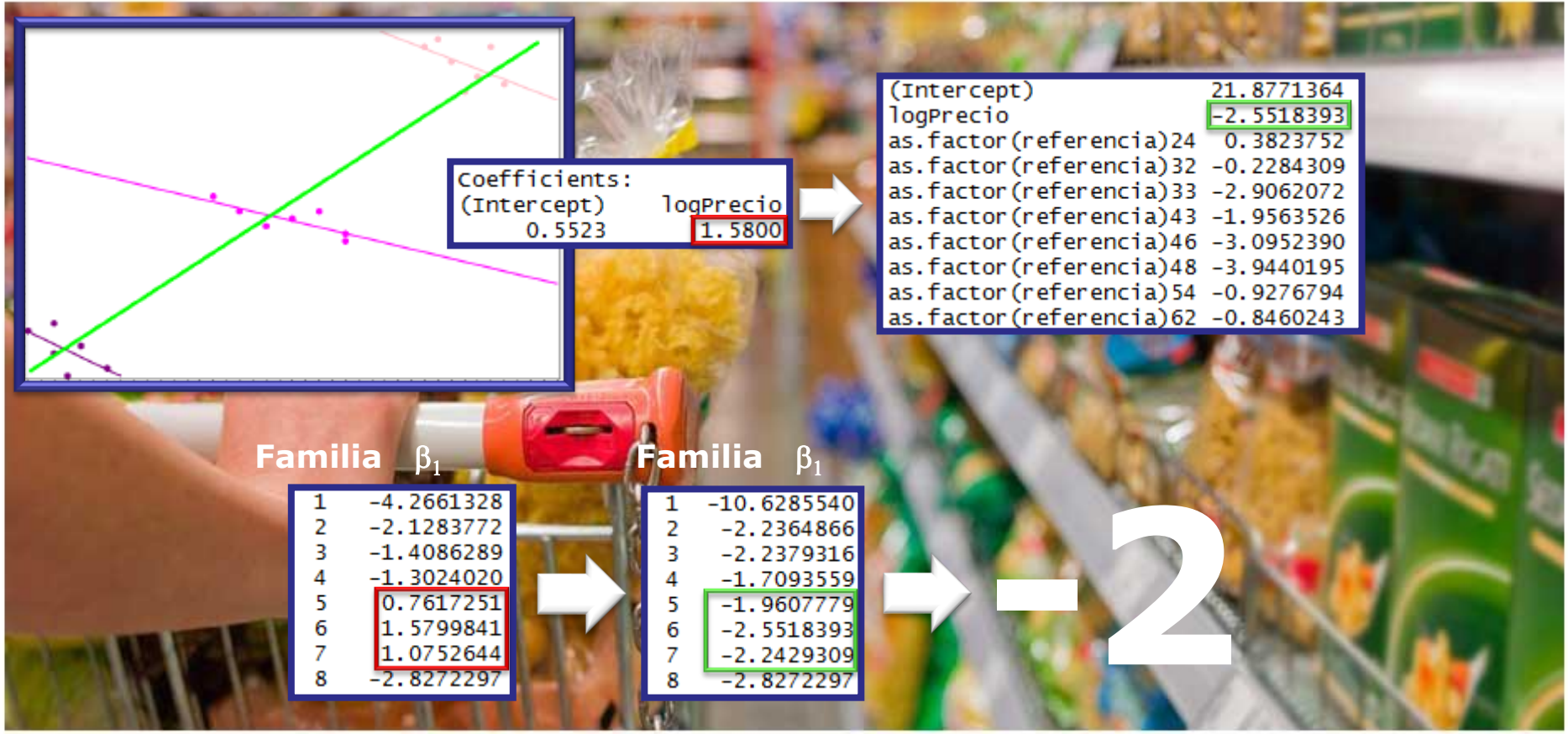
Ejemplo 1: PRICING

$$Log(Q) = \beta_0 + \beta_1 \, log(P)$$

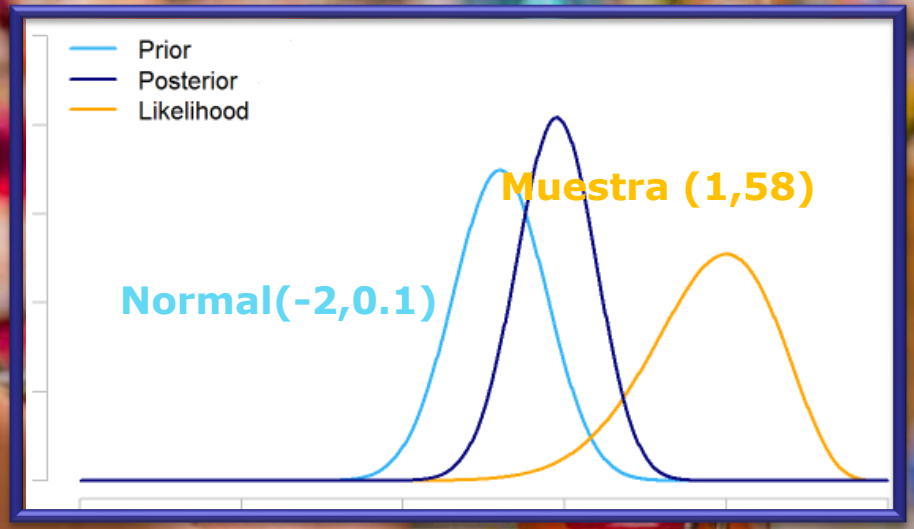
$$elasticidad = \frac{\%\Delta Q}{\%\Delta P} \sim \beta_1$$



Explicar ⇒ Sentido lógico. Buscando interpretabilidad: Datos de panel



Explicar ⇒ Sentido lógico. Buscando interpretabilidad: Regresión Bayes



El conocimiento a priori que se tiene de la relación entre las variables puede facilitar la interpretabilidad del resultado

```
Coefficients:
(Intercept)     logPrecio
     0.5523        1.5800
```

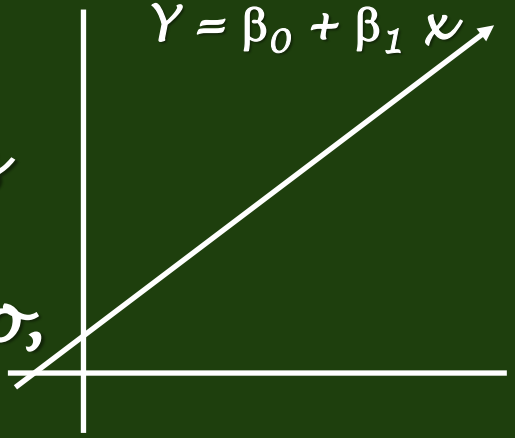
```
             Estimate
Intercept       17.25
logPrecio       -1.92
```

```
regresionBayesianaPrior <- brm(logVentas~logPrecio,
                               data = subset(datosRetailsFY,grupo==6),
                               prior = prior(normal(-2,0.1)))
```

Modelización tradicional: Regresión lineal simple

Ejemplo 2

$Y = f(x) = \beta_0 + \beta_1 x + \varepsilon$

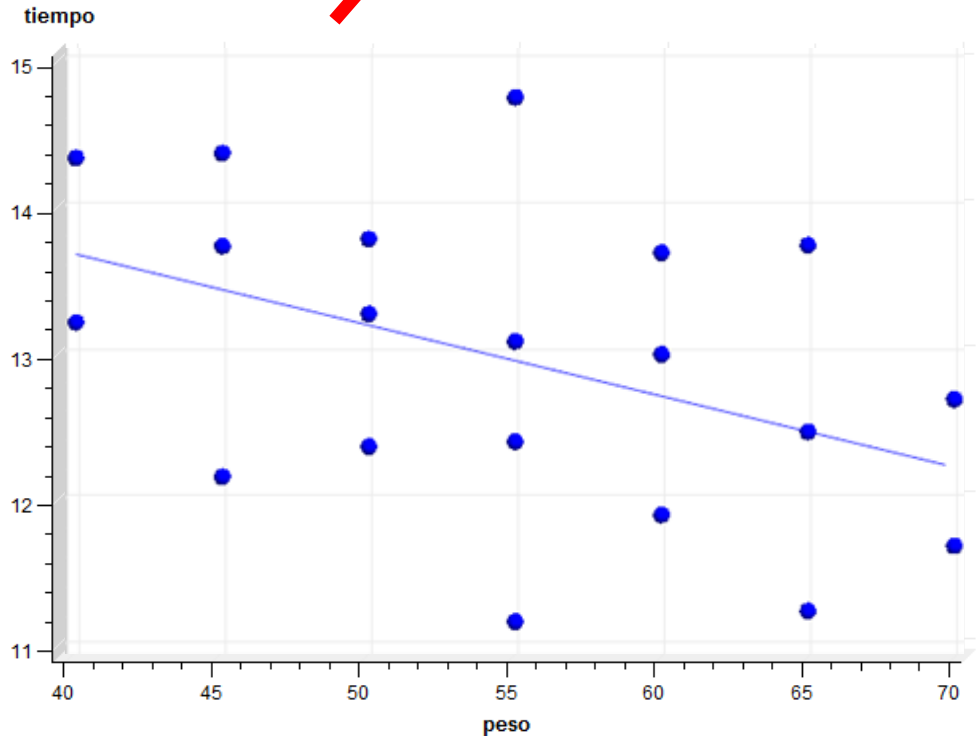


Y = tiempo en correr 100 metros

x = peso

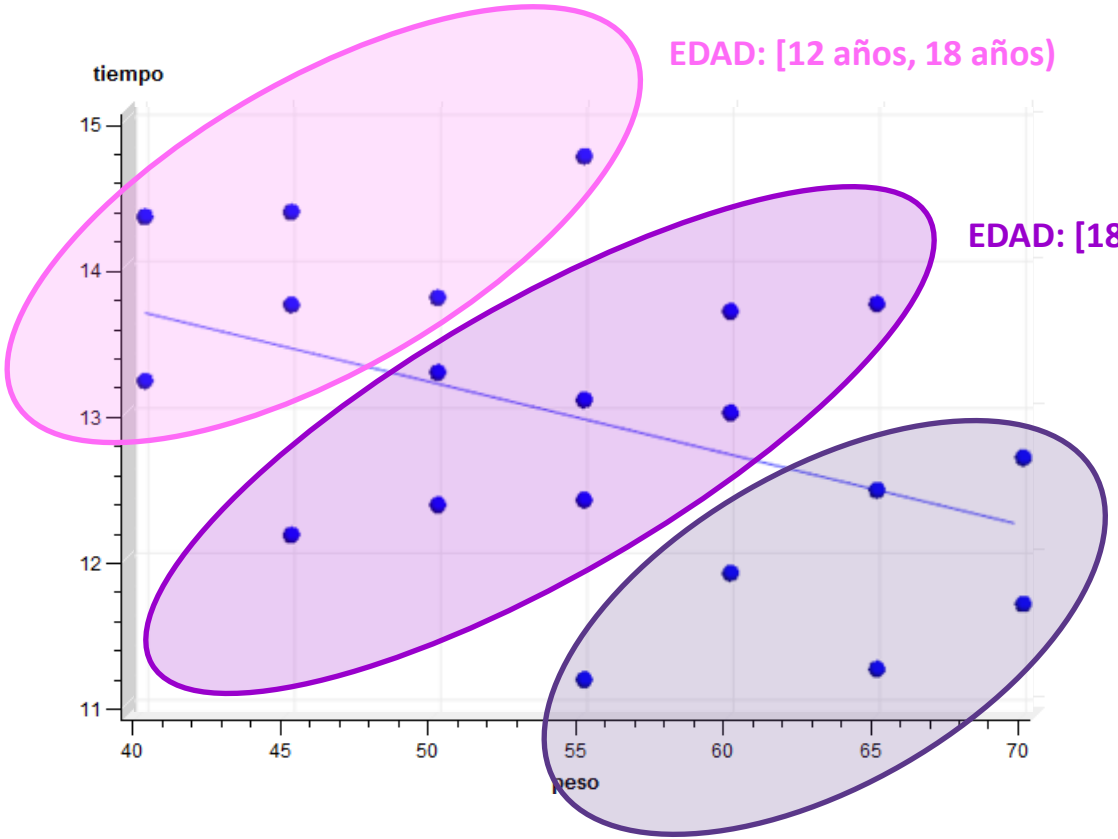
β_1 = por cada kg de más de peso, ¿cuántos segundos más se tarda en correrlos?

Explicar ⇏ Sentido lógico. Fenómeno de confusión



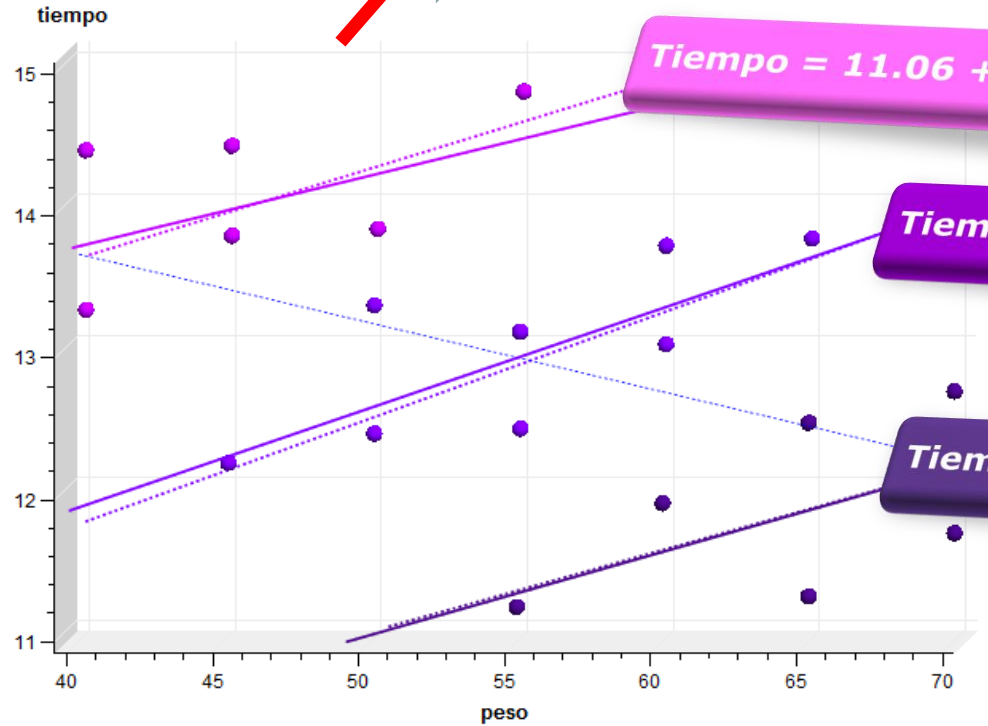
Tiempo = 15.61 - 0.05 x Peso

Explicar ⇏ Sentido lógico. Fenómeno de confusión



Incluir una variable *dummy* asociada a cada persona no parece buena opción

Explicar ⇏ Sentido lógico. Fenómeno de confusión



Tiempo = 11.06 + 0.06 * Peso

Tiempo = 8,75 + 0.07 * Peso

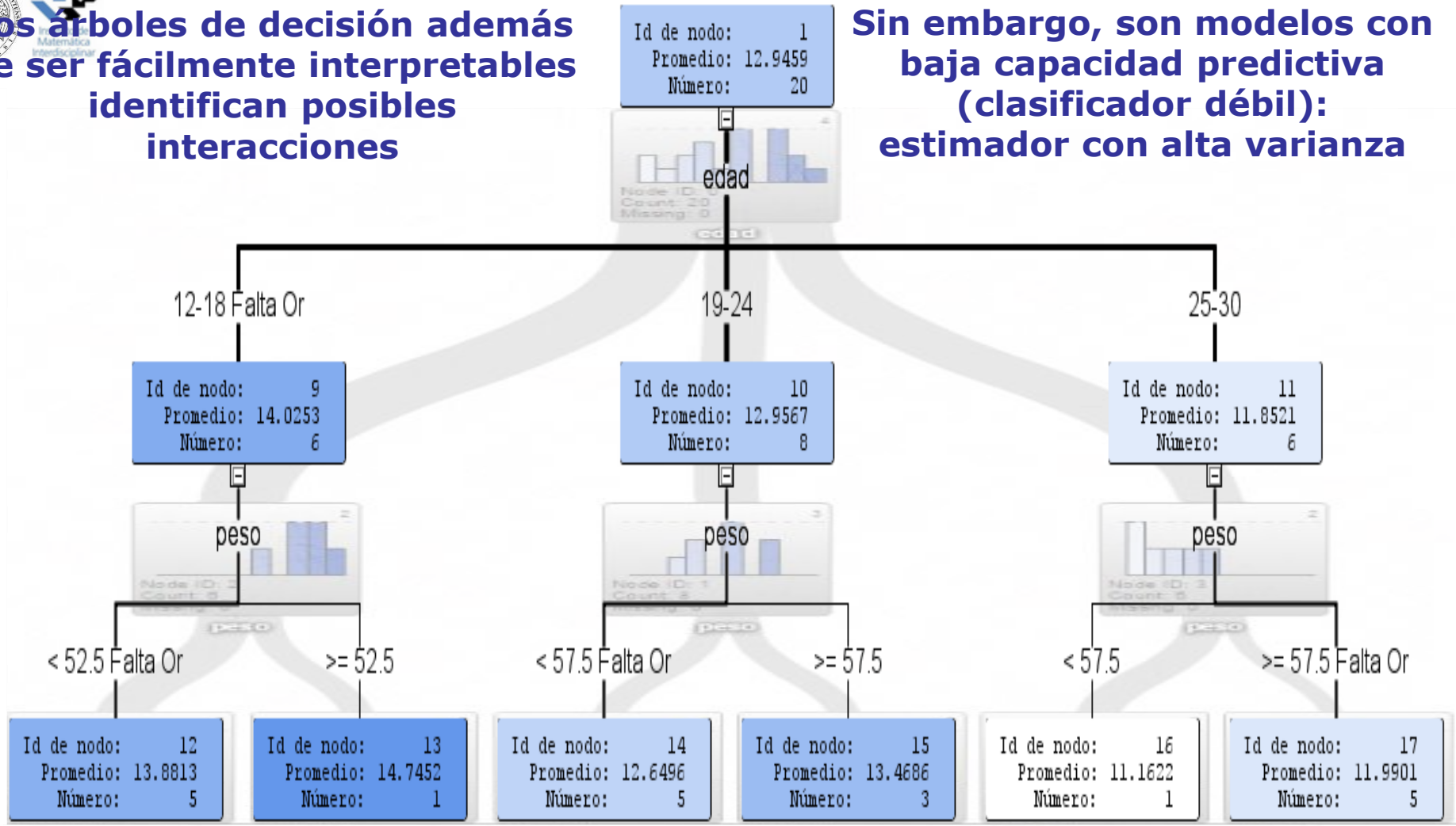
Tiempo = 8,10 + 0.06 * Peso

Tiempo = 8.09 + 0.05*Peso
+ 3.64 *I{12 ≤EDAD<18}
+ 0.99 *I{18 ≤ EDAD<25}
- 0.008 * Peso *I{12 ≤ EDAD<18}
+ 0.011 * Peso *I{18 ≤ EDAD<25}

Plantear la interacción adecuada permite explicar bien
¿Y si tengo cientos de variables?
¿Qué interacciones contemplo?

Los árboles de decisión además de ser fácilmente interpretables identifican posibles interacciones

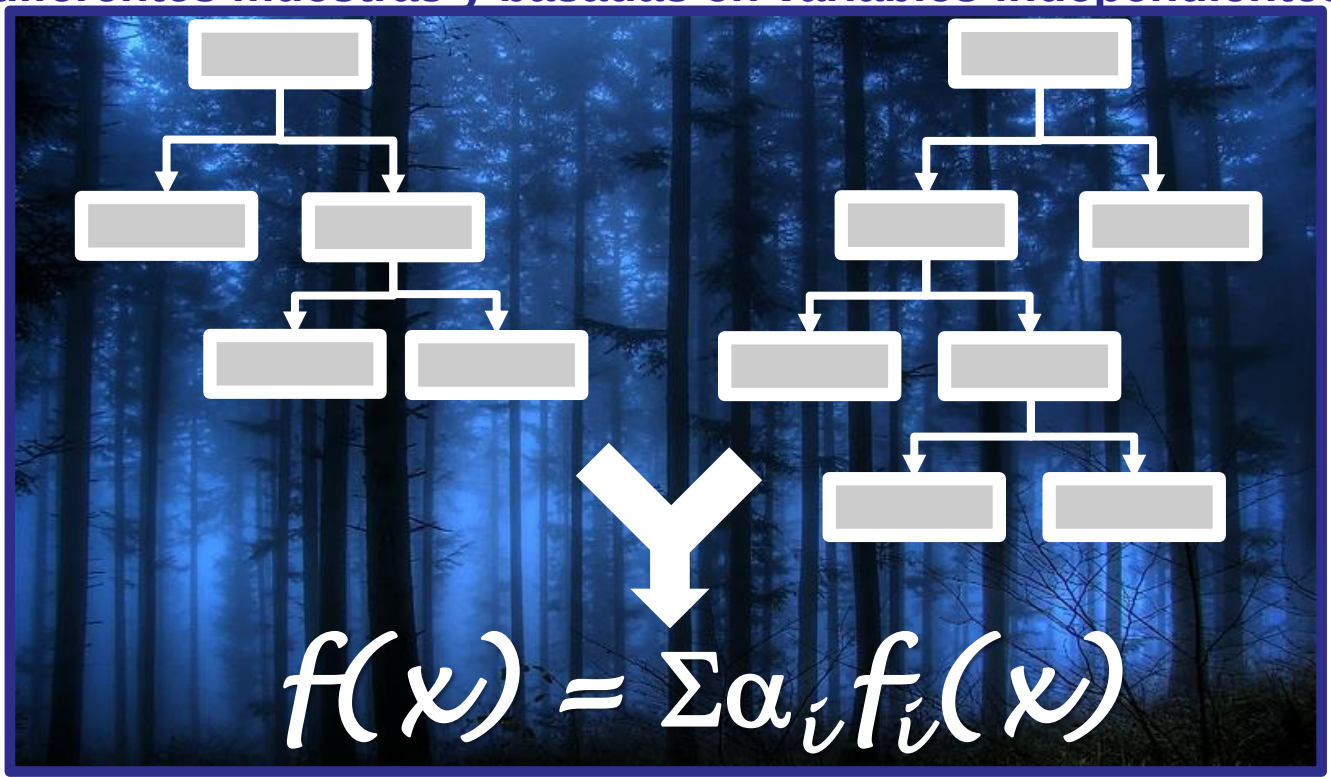
Sin embargo, son modelos con baja capacidad predictiva (clasificador débil): estimador con alta varianza



Leo Breiman
Random Forest

"Si se dispone de $X_1, \ldots, X_n$ independientes, cada una de ellas con varianza σ^2, su media $\bar{X}$ tiene varianza $\frac{\sigma^2}{n}$"

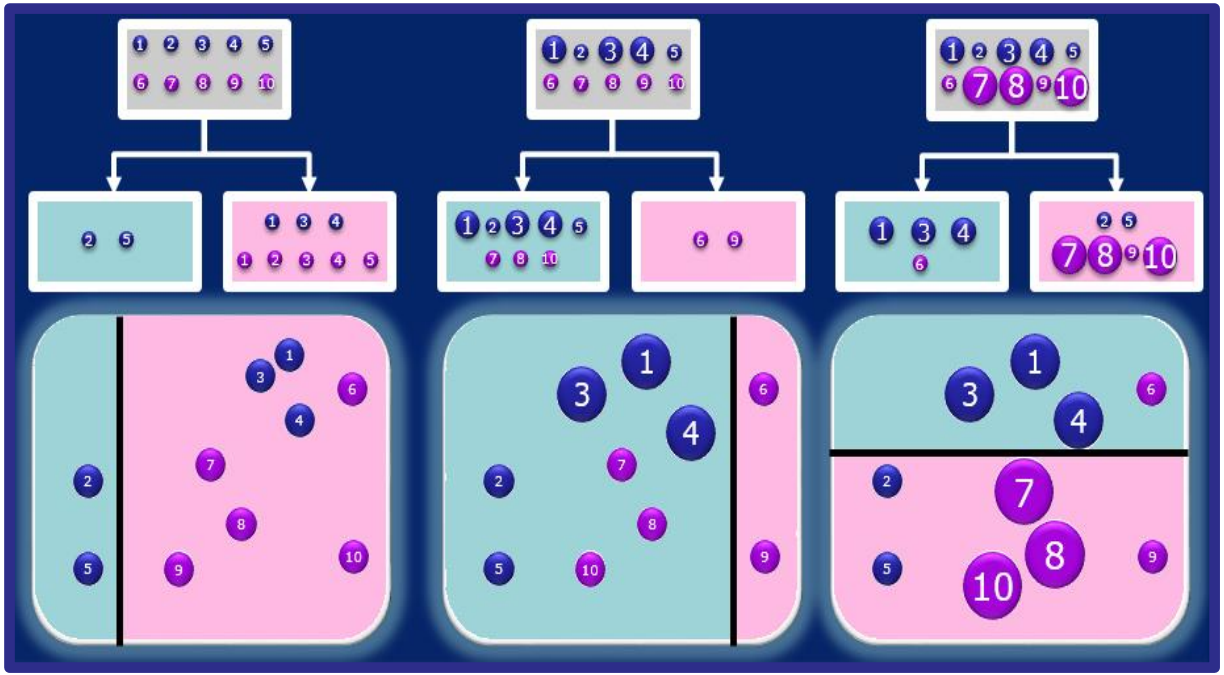
Promediar predicciones asociadas a diferentes muestras y basadas en variables independientes



¿Puede un conjunto de clasificadores débiles crear un clasificador robusto?

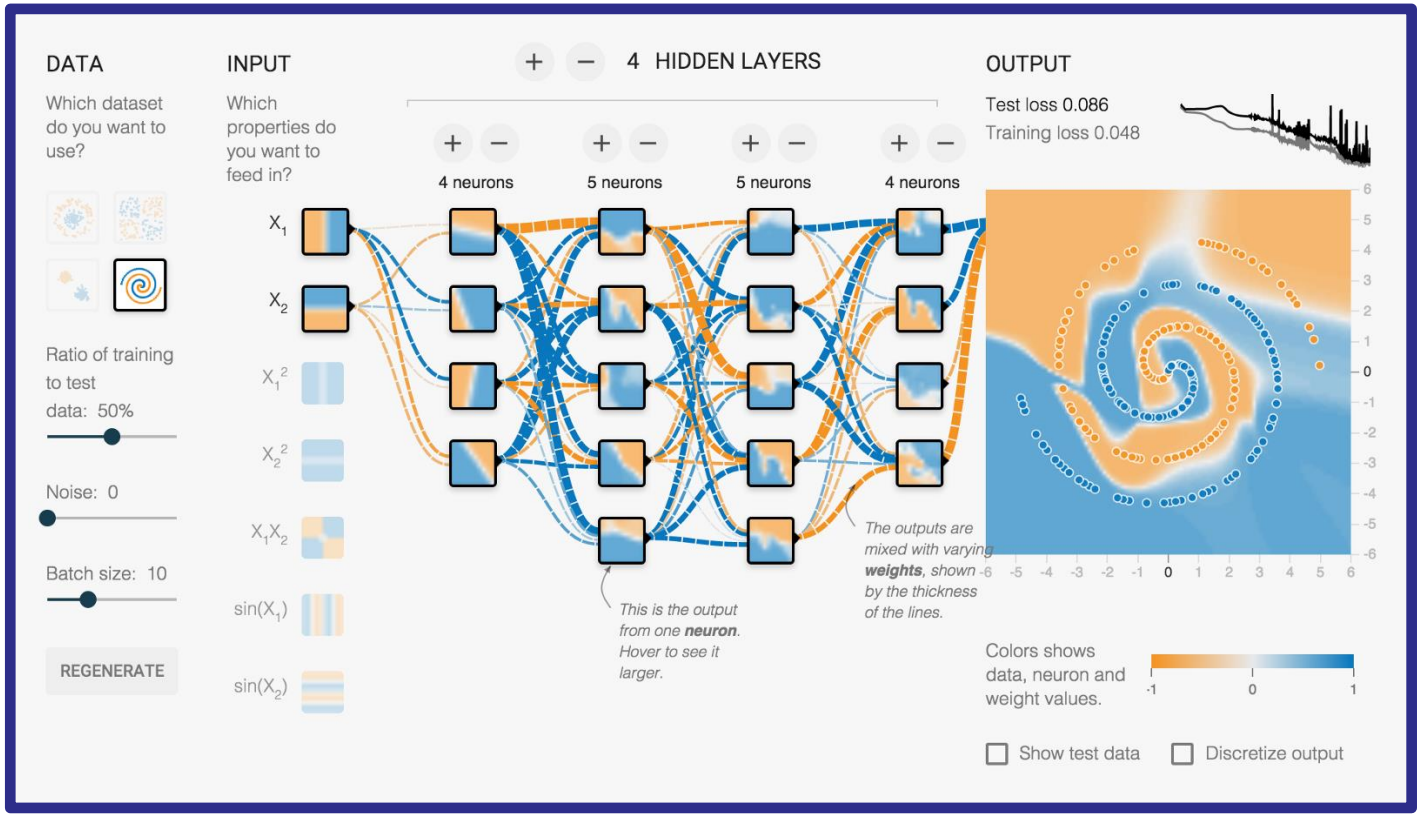
Robert Schapire *Yoav Freund*

Adaboost
Premio Gödel 2003



Los modelos tipo "*boosting*" se han convertido en un referente dentro de la modelización automática (*Machine Learning*) por su alto nivel de predictibilidad.

Las redes neuronales es otra de las estrategias de modelización más competitivas dentro del ámbito del *Machine Learning*



http://playground.tensorflow.org

¿Hasta qué punto sacrificar la capacidad de interpretación de un modelo por su capacidad predictiva?

Casos de éxito de estrategias tipo "*boosting*"

Net Promoter Score

ESPAÑA (UCM): 56,5%
EL SEGUNDO: 52,6%

Predecir el grado de satisfacción de un cliente

Regresión: 52% ➡ **Boosting: 54%**

Fraude energético

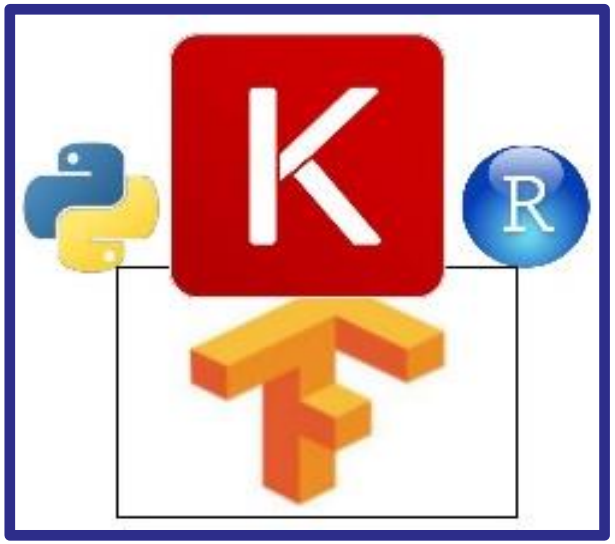
INNOVA-TSN: 80%
EL SEGUNDO: 60%

Seleccionar clientes sospechosos de manipular su contador eléctrico

Regresión: 40% ➡ **Boosting: 80%**

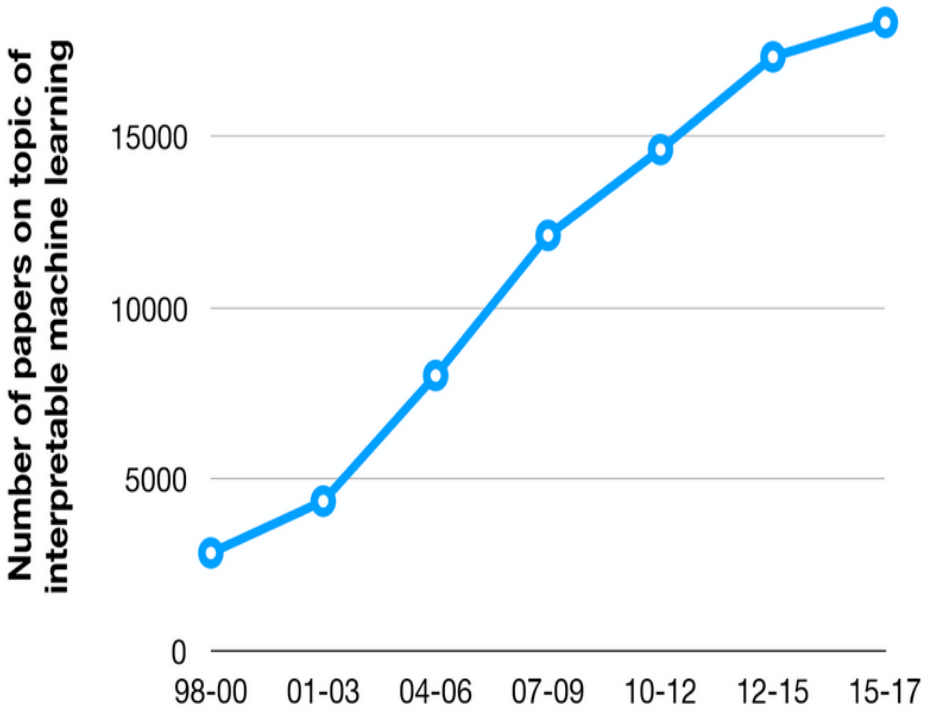
La alta capacidad predictiva de estos modelos ha justificado el desarrollo de softwares que implementan eficientemente este tipo de técnicas:

H_2O es un referente claro dentro del ámbito del Aprendizaje Automático
KERAS es un referente claro dentro de un ámbito más específico: *Deep Learning*



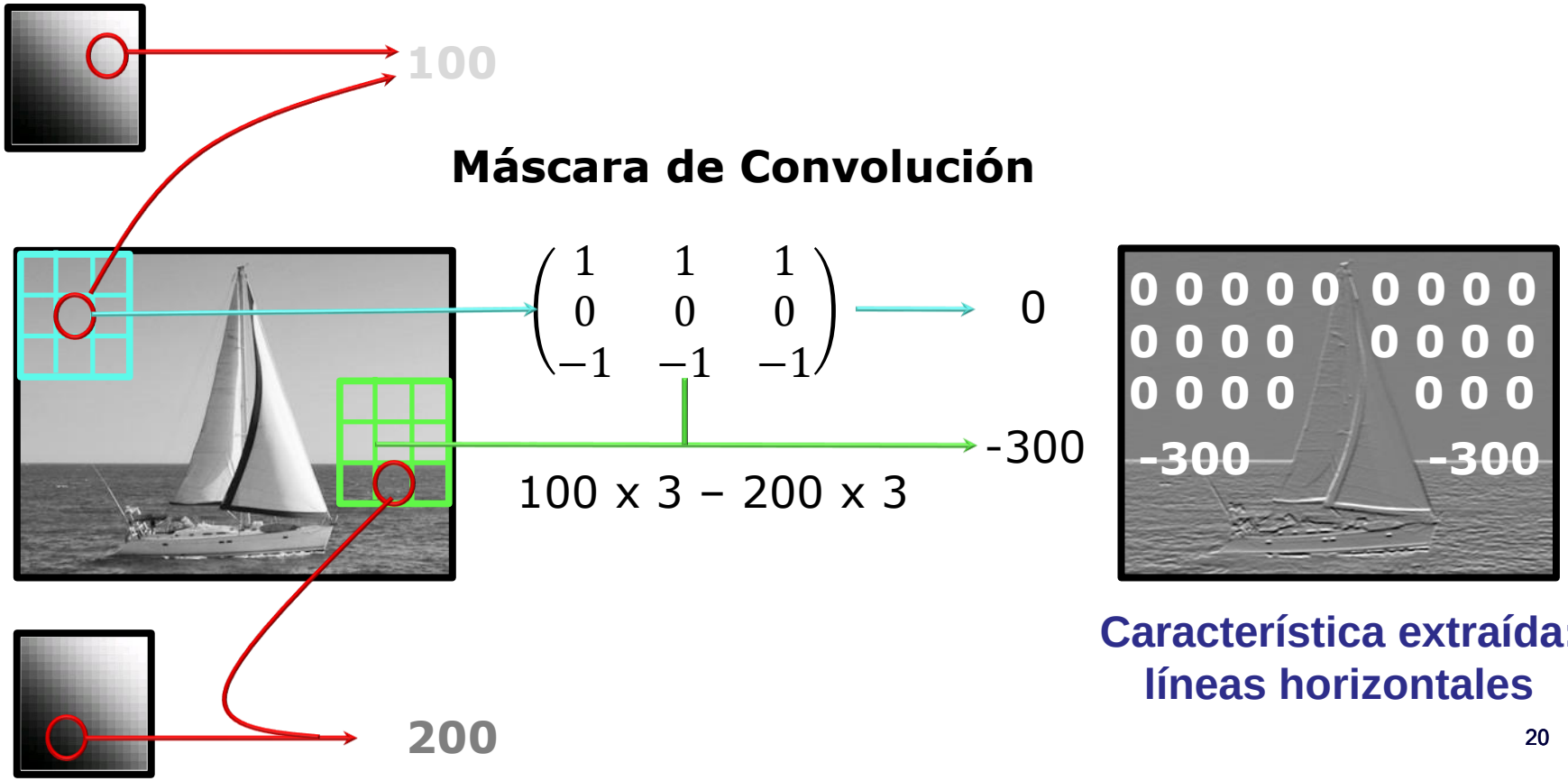
https://www.youtube.com/watch?v=wcyMBRRLmqs

En algunos ámbitos como finanzas o la salud, es preciso auditar el proceso de decisión y asegurar de que no sea discriminatorio ni viole la ley



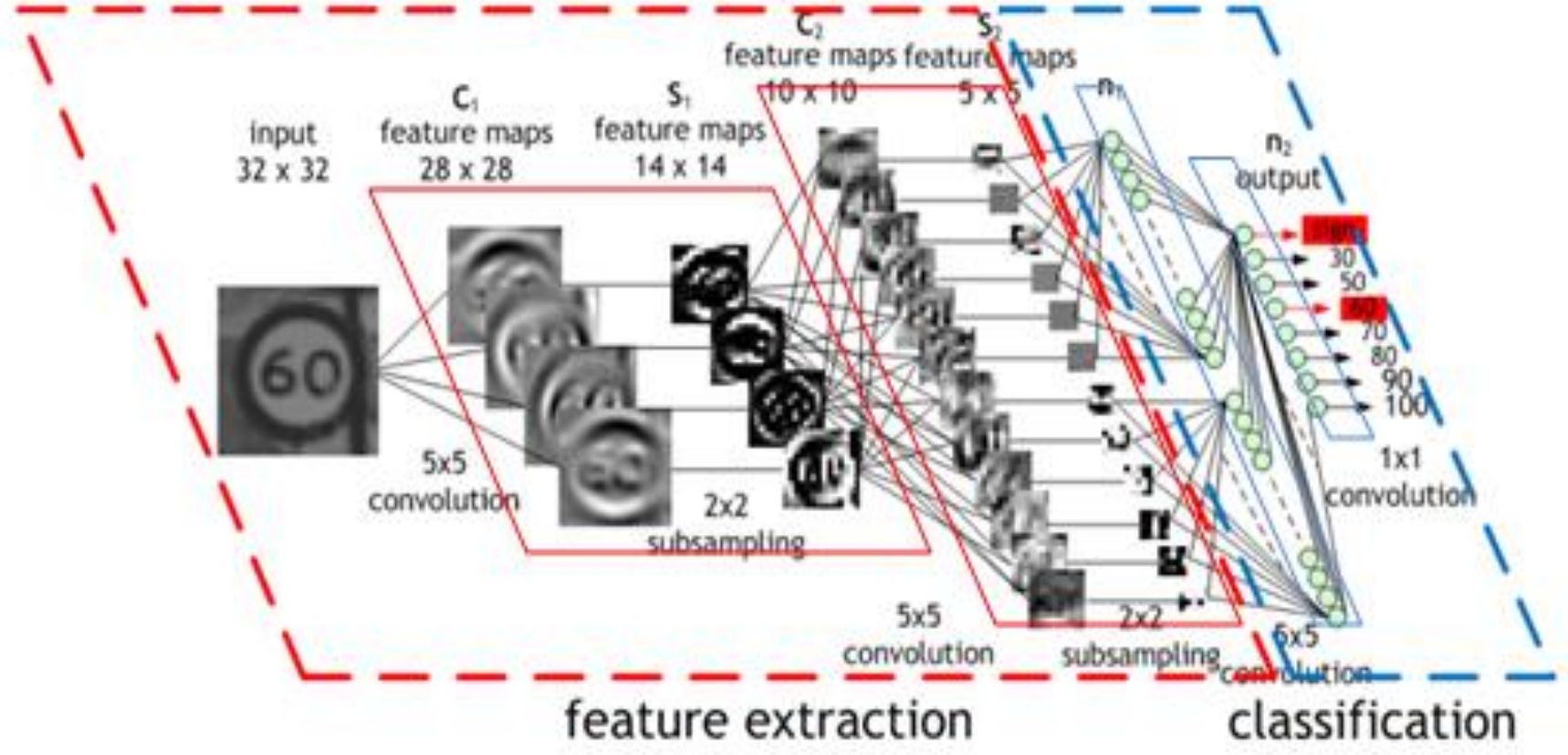
Por ello, en los últimos años han proliferado sensiblemente los algoritmos de *Machine Learning* que buscan la interpretabilidad

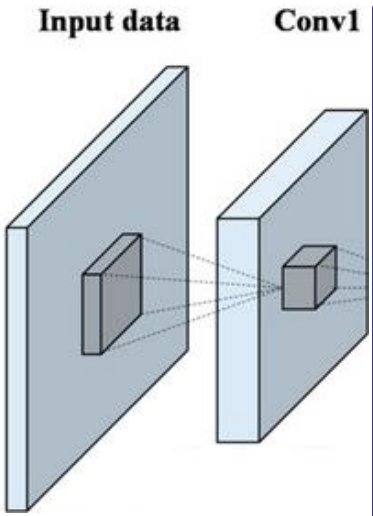
Algunos modelos, como las redes de convolución, ayudan a entender mediante la extracción de características, qué está teniendo en cuenta un modelo cuando identifica una imagen



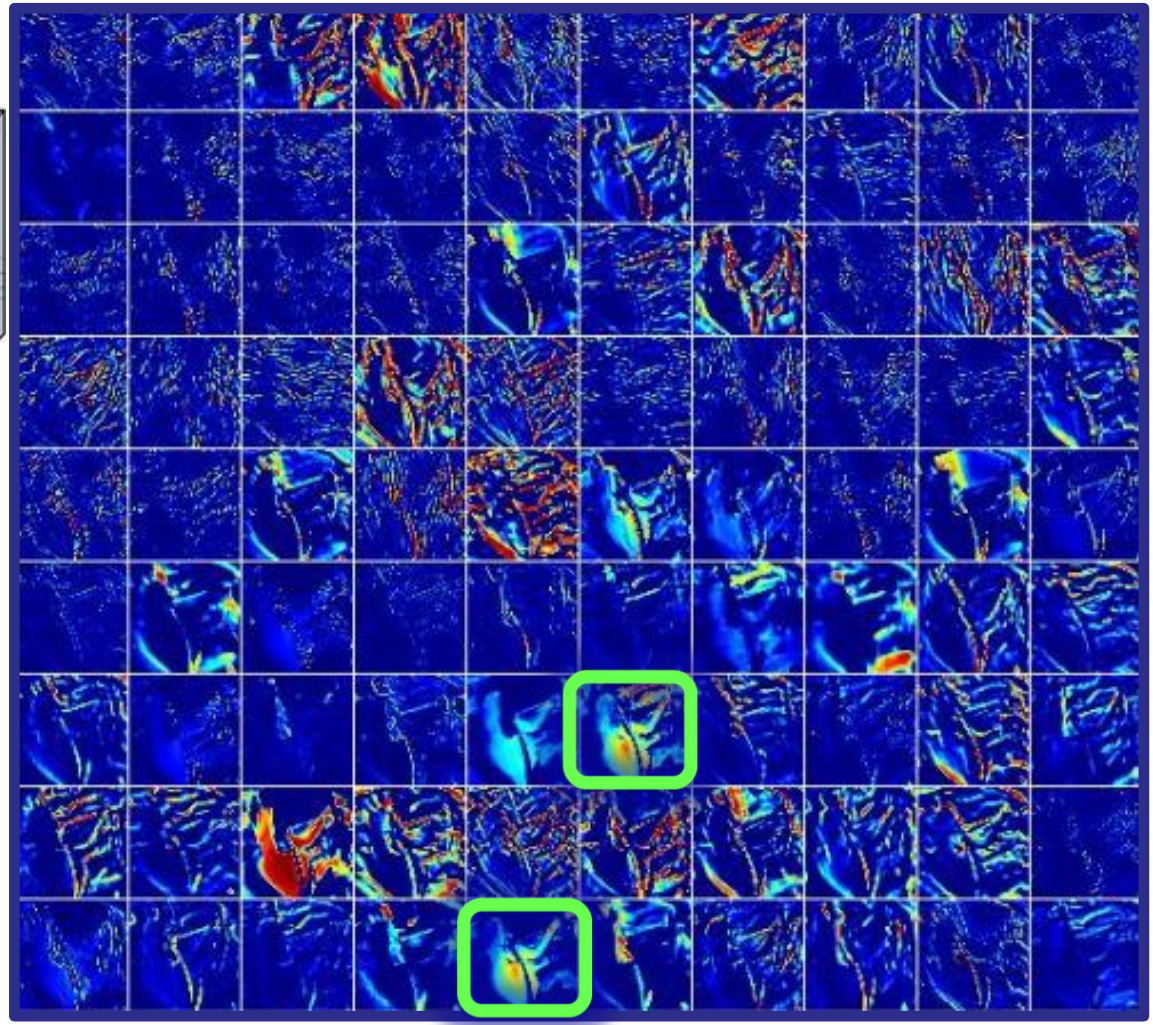
Característica extraída: líneas horizontales

Las capas de convolución permiten la extracción de características como paso previo al ajuste del modelo que lleva a cabo la predicción/clasificación





Ejemplo: Identificación mediante la imagen captada por una cámara de personas hablando por el móvil



Sin embargo, lo ideal es la búsqueda de metodologías que permitan interpretar los resultados de un modelo con independencia de la naturaleza de éste

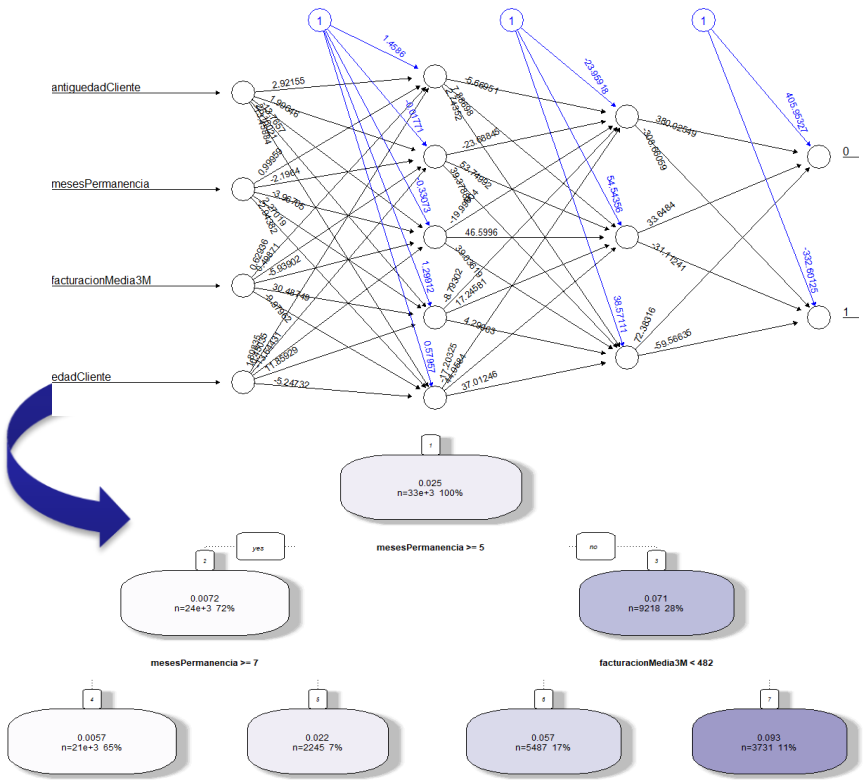
Metodologías globales: ¿Cuáles son las variables que más importancia tienen en la predicción?¿Cómo influyen?

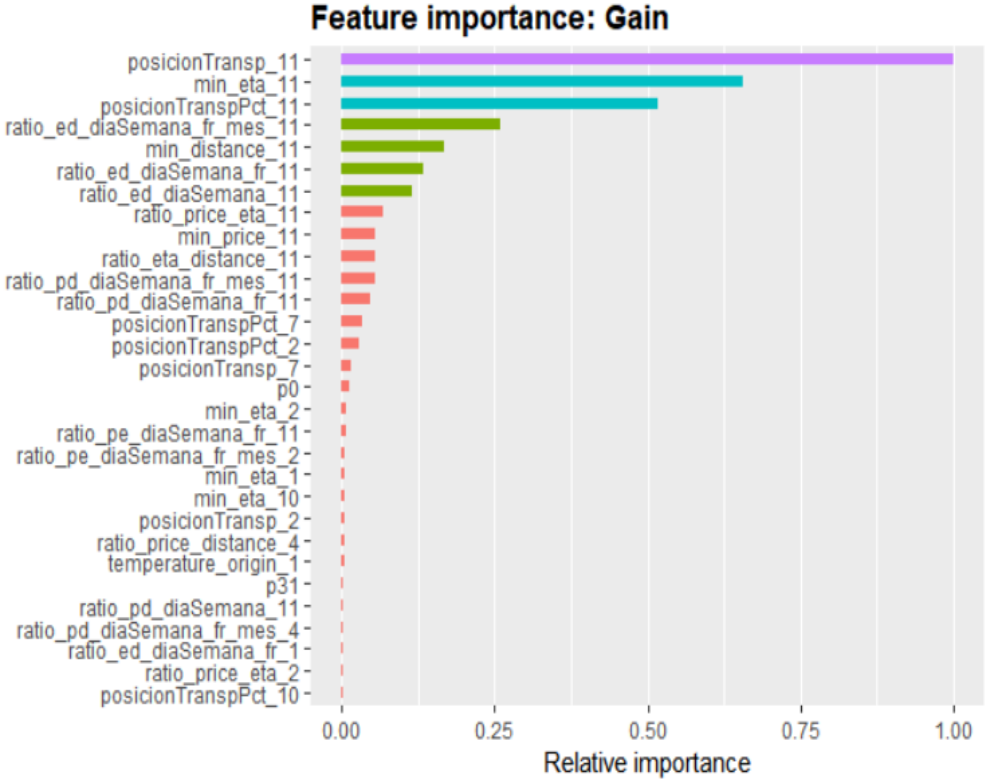
- Árbol de decisión
- Gráficos de importancia de variables

Metodologías locales: ¿Cuáles son las variables que más condicionan la predicción dada a un individuo? (ejemplo: ¿de qué sirve saber que un cliente va a dejar mi compañía si no conozco la razón (variable) que lo justifica?

- LIME: *Local Interpretable Model-agnostic Explanations*
- DALEX: *Descriptive mAchine Learning Explanations*
- IML: *Interpretable Machine Learning*

Metodologías globales





En un problema de clasificación binaria, un árbol de regresión permitiría explicar la probabilidad predicha por un modelo

Un gráfico de importancia de variables permite ver en cuántos de los modelos ensamblados entra cada una de las variables o su contribución

LIME: Una metodología local

Se trata de una metodología que busca justificar, registro a registro, cuál es la variable que tiene mayor contribución al valor predicho para él

Para ello, se procede de la siguiente manera:

1. **Se selecciona un registro**
2. **Se generan registros nuevos en un entorno suyo mediante variaciones aleatorias del valor de las explicativas asociadas a dicho registro, dando más peso a las más similares**
3. **Se calcula la predicción asociada a los registros**
4. **Se ajusta un modelo sencillo e interpretable para separar los registros predichos en una y otra clase**

Registros simulados:

1 "Me gusta película" ➡ predicción 1
2 "Me esta película" ➡ predicción 2
3 "Me película" ➡ predicción 3
4 "Me gusta" ➡ predicción 4

Ajustar un modelo interpretable (árbol) para explicar los registros en función de las palabras

Conclusiones

En algunas problemáticas, las predicciones proporcionadas por un modelo "*black box*" pueden ser sensiblemente mejores que las de uno explicable
Sin embargo, las decisiones tomadas por dicho modelo no siempre se entienden bien

Es importante poder interpretar dichas predicciones para poder tomar decisiones justificadas


¡Muchas
Gracias!