Machine Learning avanzado con algoritmos híbridos

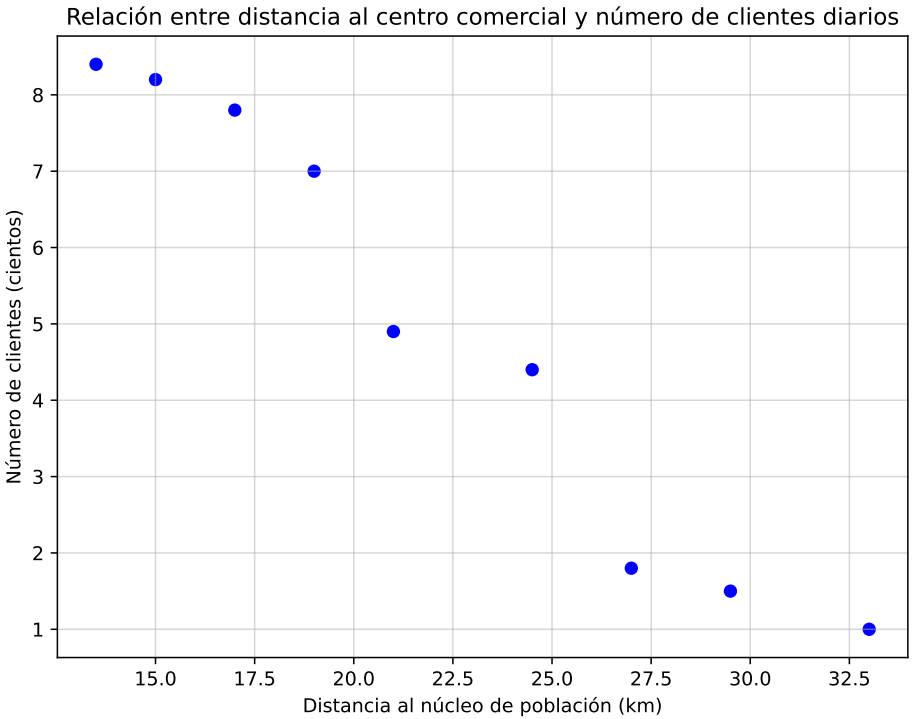
Manuel García Plaza

Universidad de Alicante

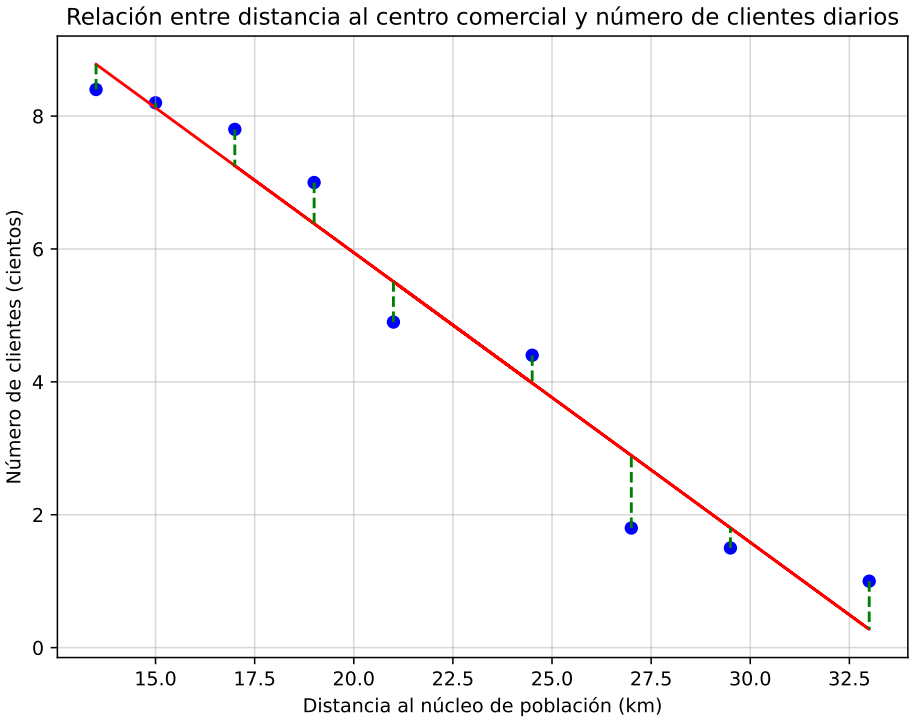
14 de junio de 2024

Dos tipos de problemas en los casos de uso:

- De regresión: predecir un valor numérico continuo.
  - Predicción de demanda eléctrica.
  - Predicción de temperaturas críticas en superconductores.
  - Estimación de precios de viviendas.
- De clasificación: asignar una categoría.
  - Detección de cáncer de mama.
  - Localización de defectos de software.
  - Identificación de clientes reclamantes de un seguro.



Tendencia lineal en la nube de puntos.



Tendencia lineal en la nube de puntos.

Se busca la recta $y = \beta_0 + \beta_1 x$, con β_0 y β_1 tales que los errores sean lo más pequeños posible.

En general, cada punto queda determinado por la ecuación

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \ ,$$

y todo el conjunto de datos, por la ecuación matricial

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \ ,$$

o, equivalentemente,

$$y = X\beta + \varepsilon \ .$$

Se resuelve por Mínimos Cuadrados.
La cantidad a minimizar es:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) \ ,$$

respecto del vector de coeficientes β.
La expresión de la solución óptima es:

$$\hat{\beta} = (X^T X)^{-1} X^T y \ ,$$

y las estimaciones se obtienen a partir de:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \ .$$

Problema para la Regresión Lineal: las variables binarias.

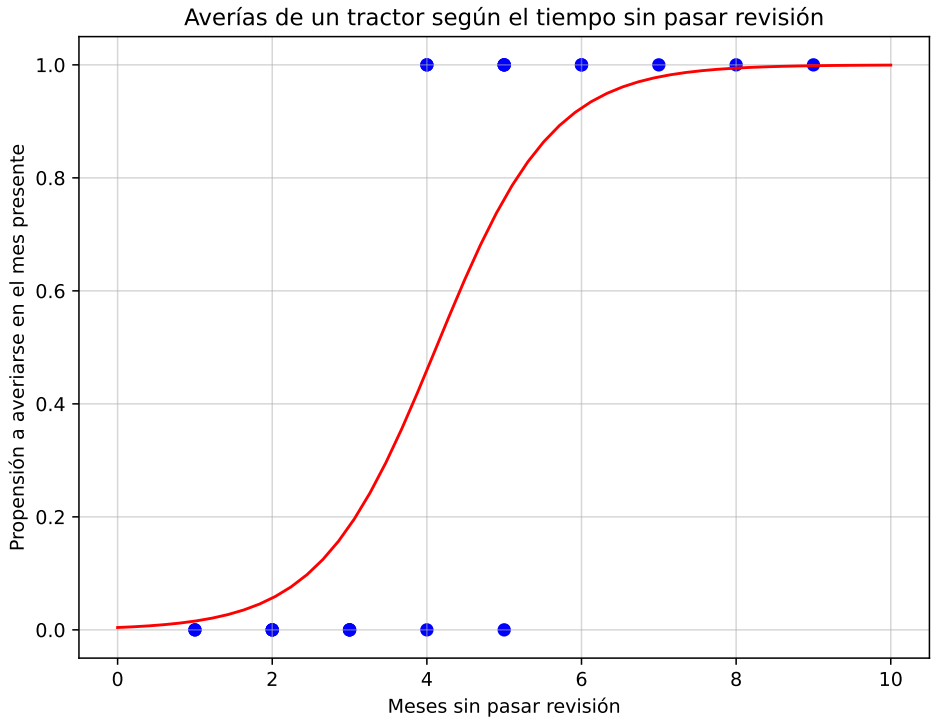
Problema para la Regresión Lineal: las variables binarias.

¿Cómo se interpreta?

Solución: transformar el modelo de Regresión Lineal en Regresión Logística para poder predecir probabilidades.

La función usada para la conversión es la sigmoide:

$$\sigma(x) = \frac{e^x}{1 + e^x} \ .$$



El modelo resultante es:

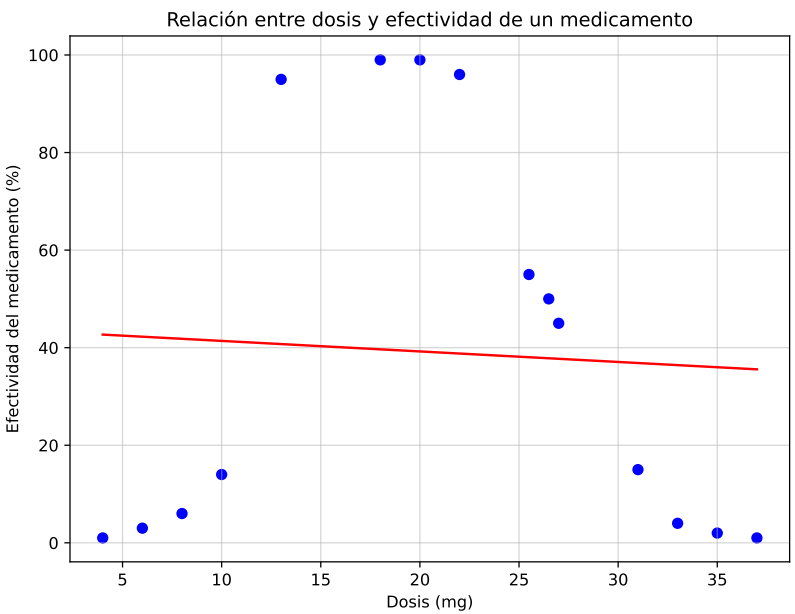
$$p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}} \ .$$

El vector de coeficientes β se obtiene mediante el método de Máxima Verosimilitud.
Se maximiza

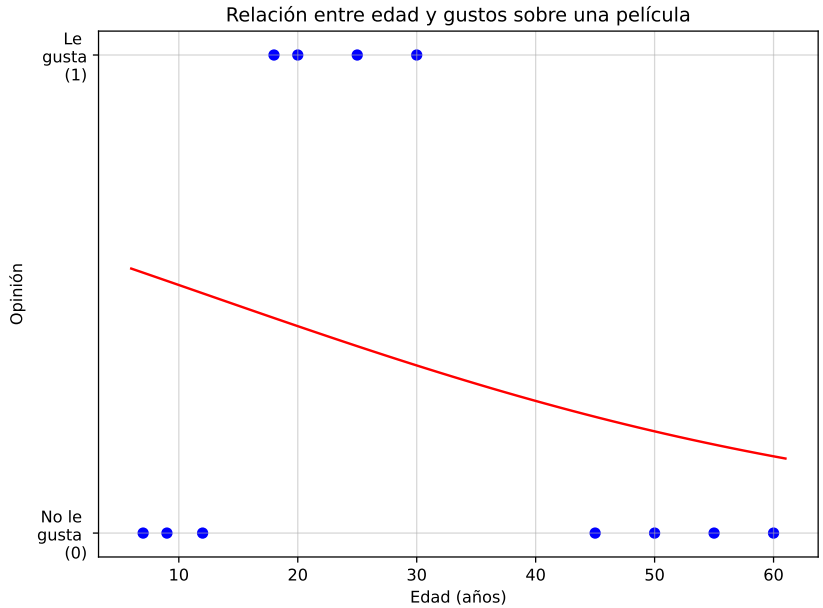
$$\mathcal{L}(p|y_1,\ldots,y_n) = \prod_{i=1}^{n} f_p(y_i) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \ .$$

Problema para los métodos de Regresión: las relaciones no lineales.

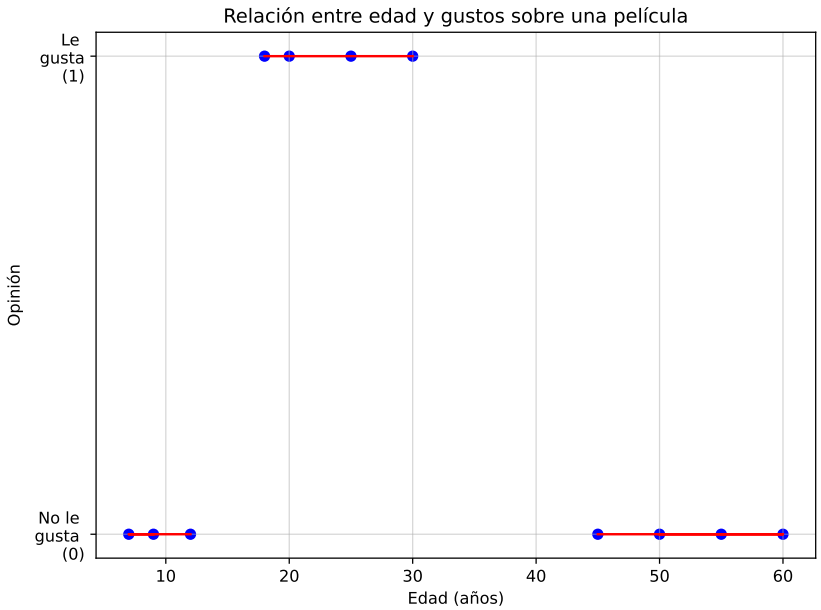
Caso de regresión:

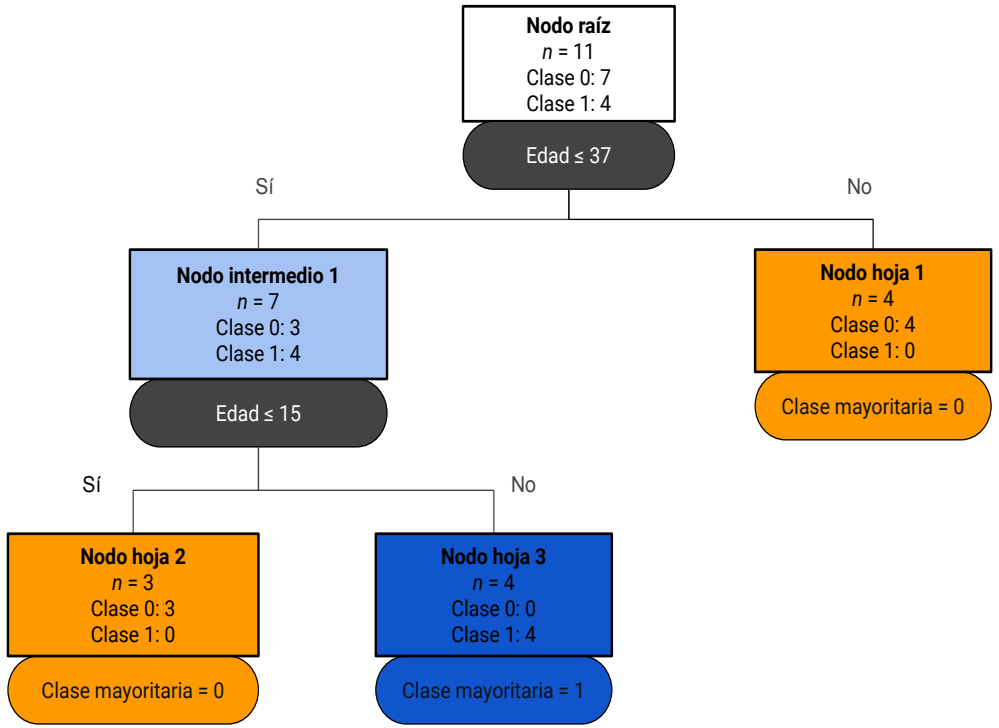


Caso de clasificación:

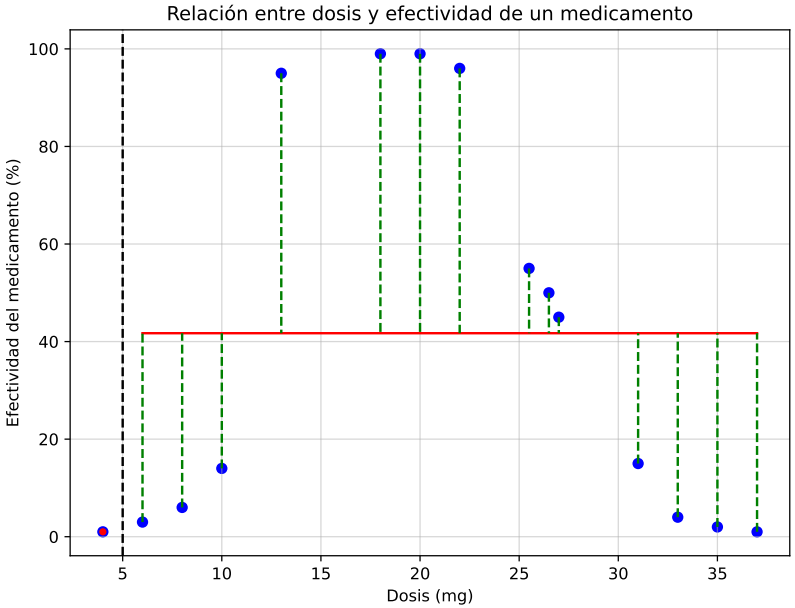


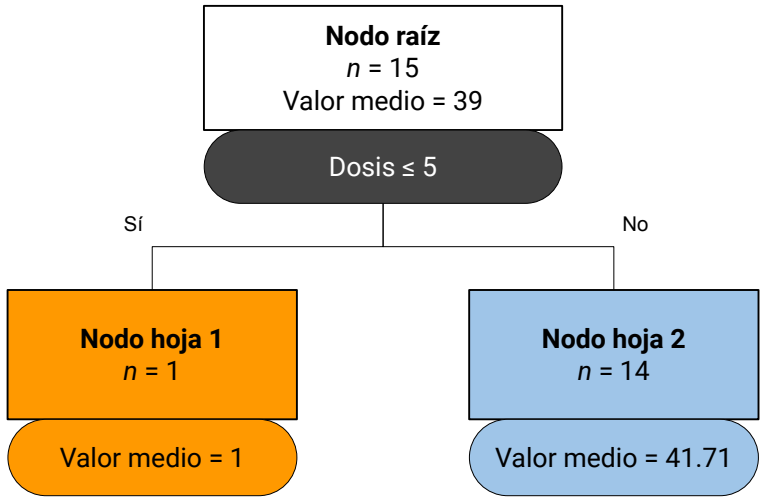
Solución: Árboles de Decisión (CART).





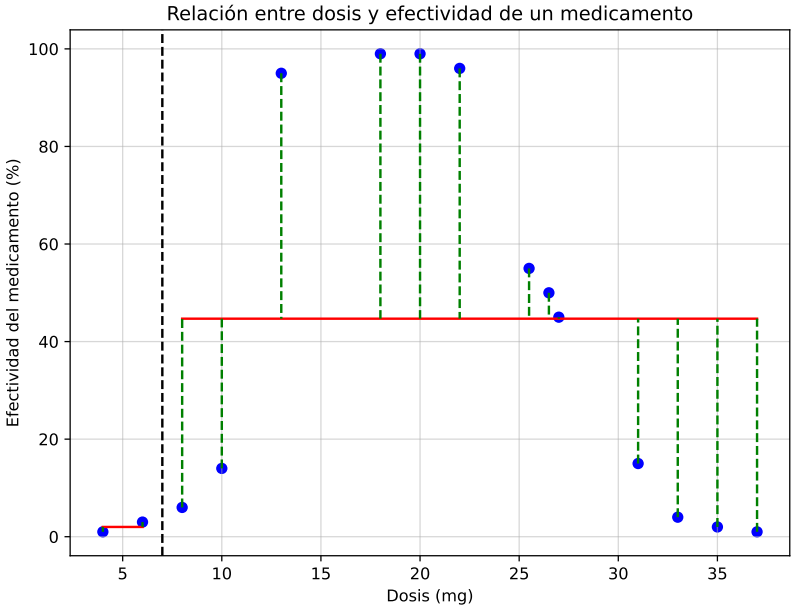
¿Cómo funcionan?

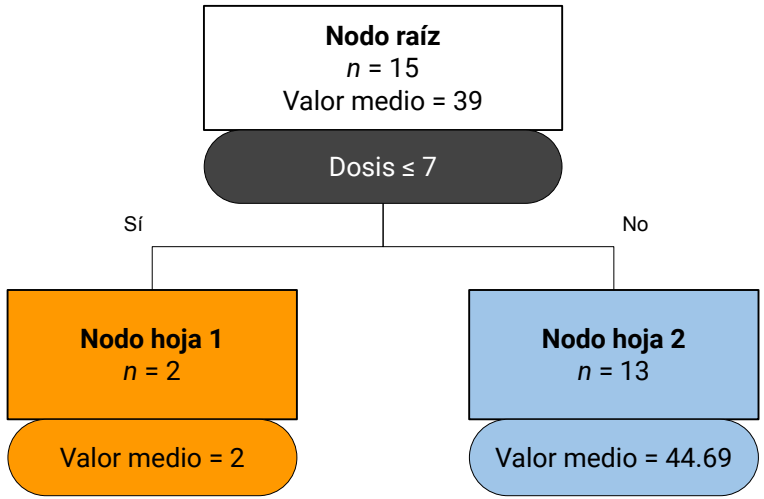




$$\sum_{i=1}^{15} \varepsilon_i^2 = 21518.86$$

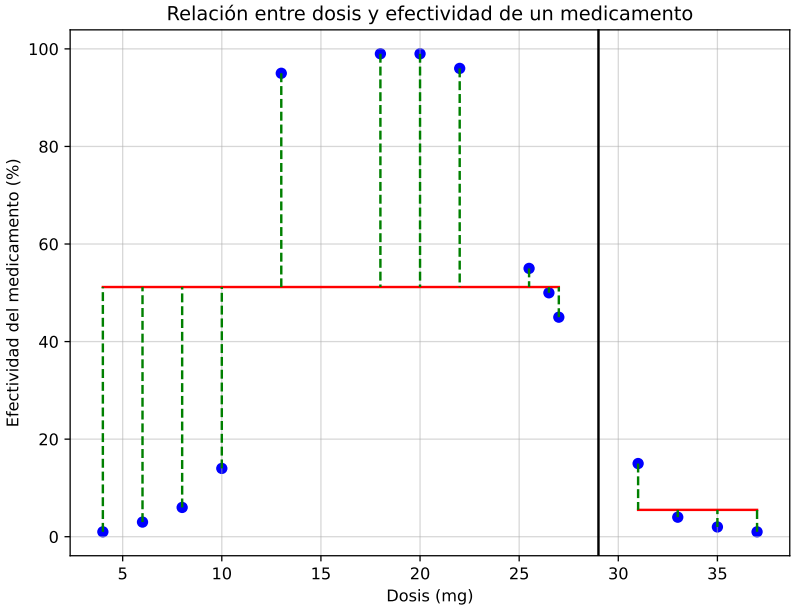
¿Cómo funcionan?

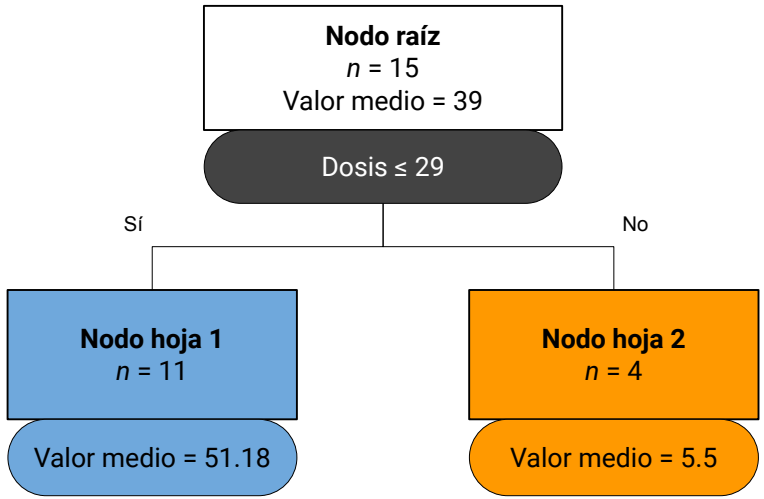




$$\sum_{i=1}^{15} \varepsilon_i^2 = 19906.77$$

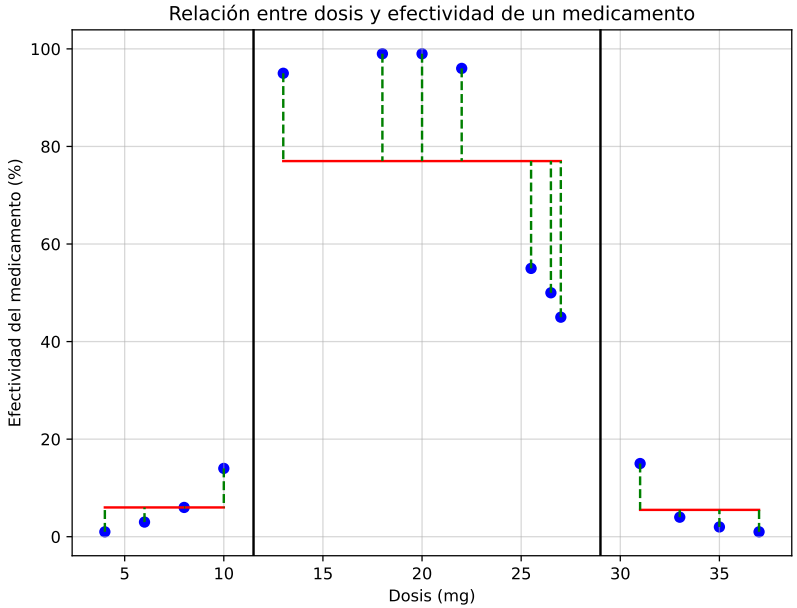
¿Cómo funcionan?

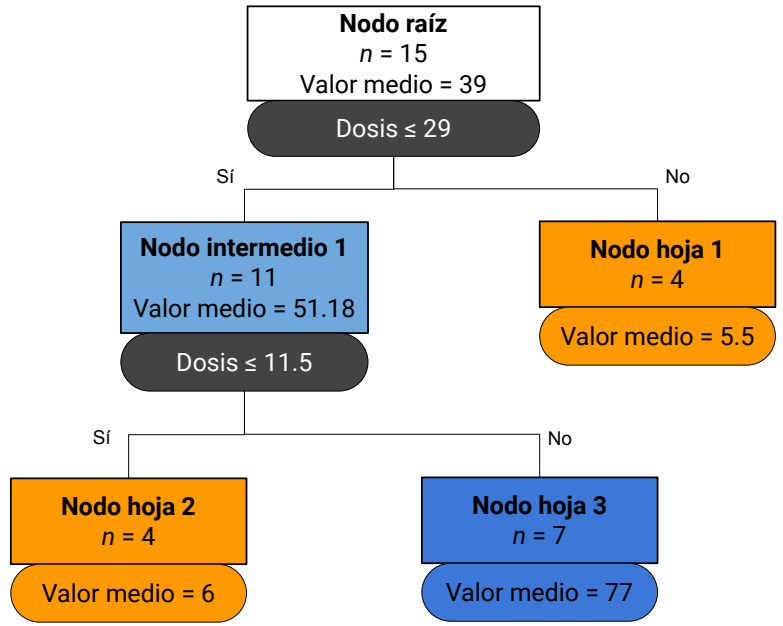




$$\sum_{i=1}^{15} \varepsilon_i^2 = 16819.64$$

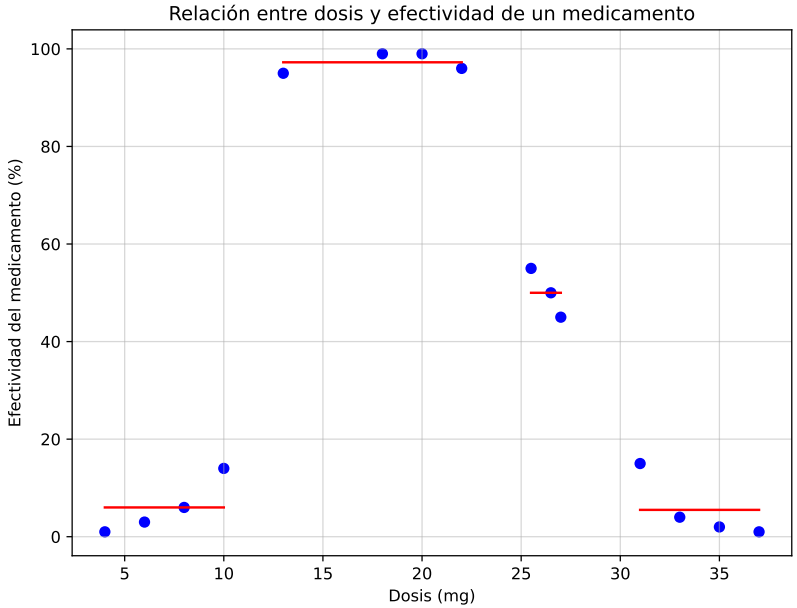
¿Cómo funcionan?

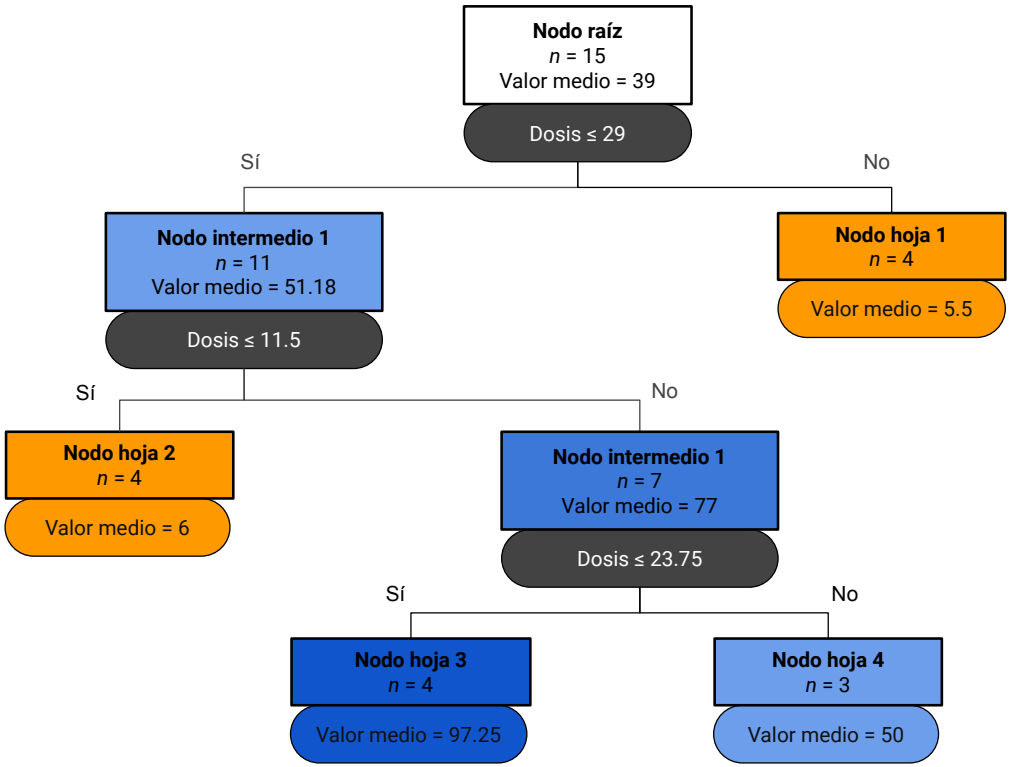




$$\sum_{i=1}^{15} \varepsilon_i^2 = 4113$$

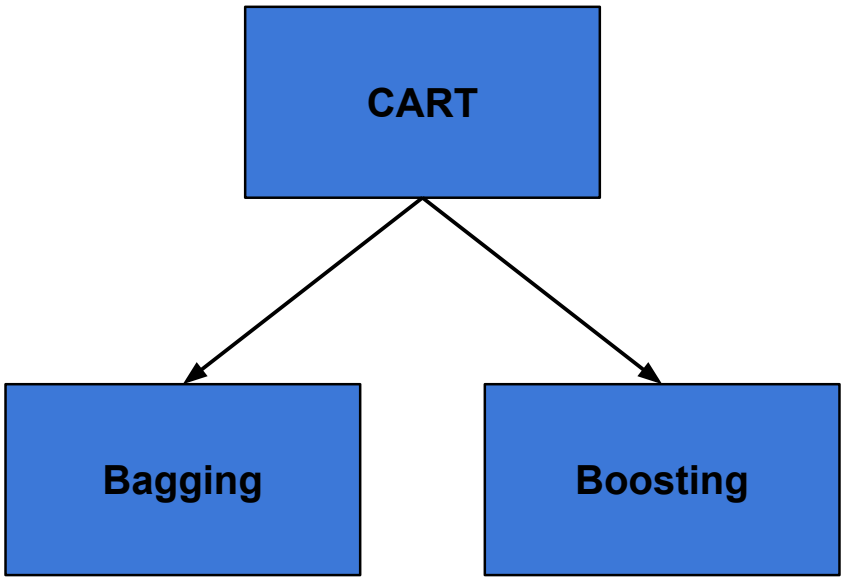
¿Cómo funcionan?



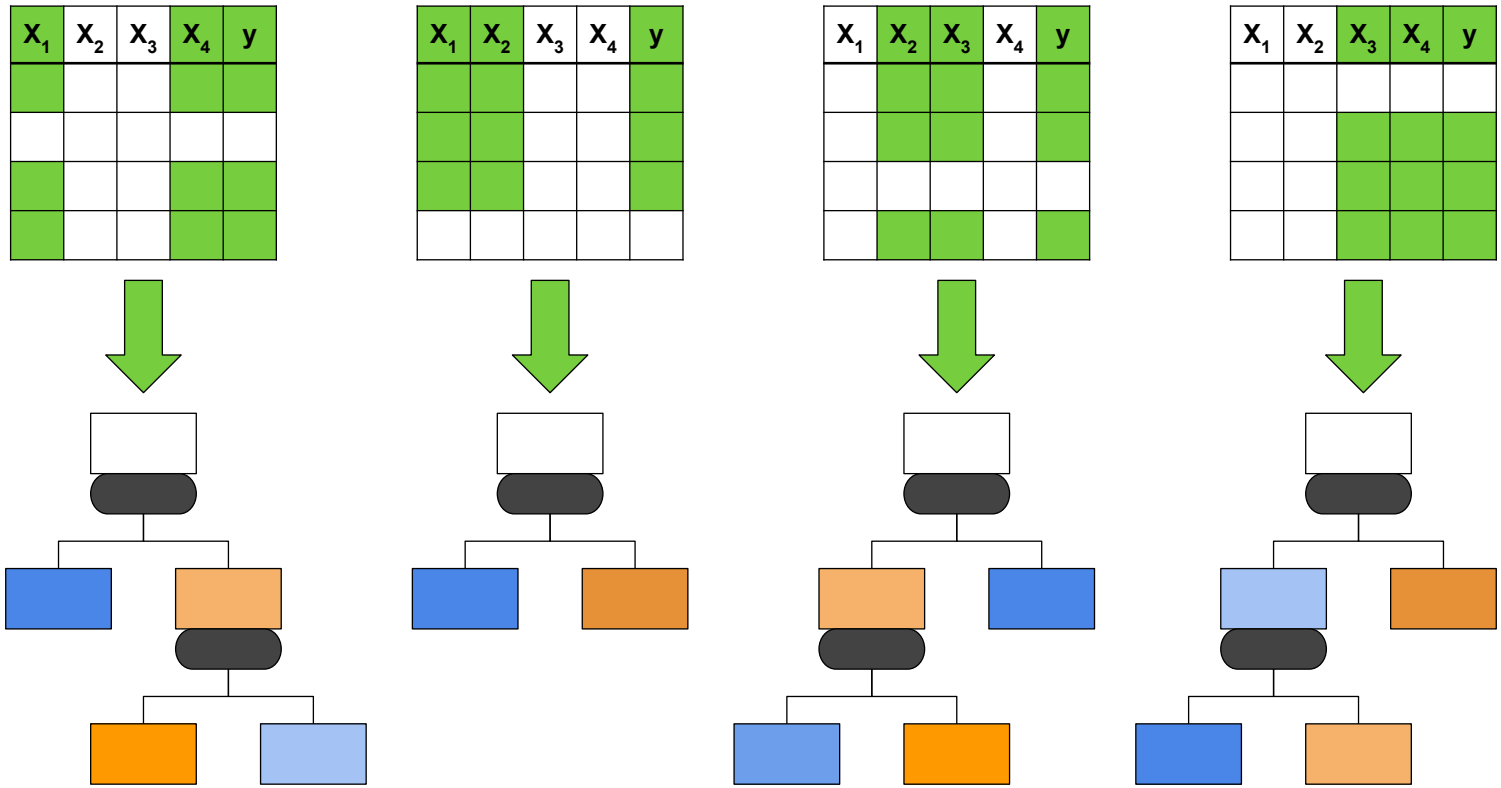


$$\sum_{i=1}^{15} \varepsilon_i^2 = 285.75$$

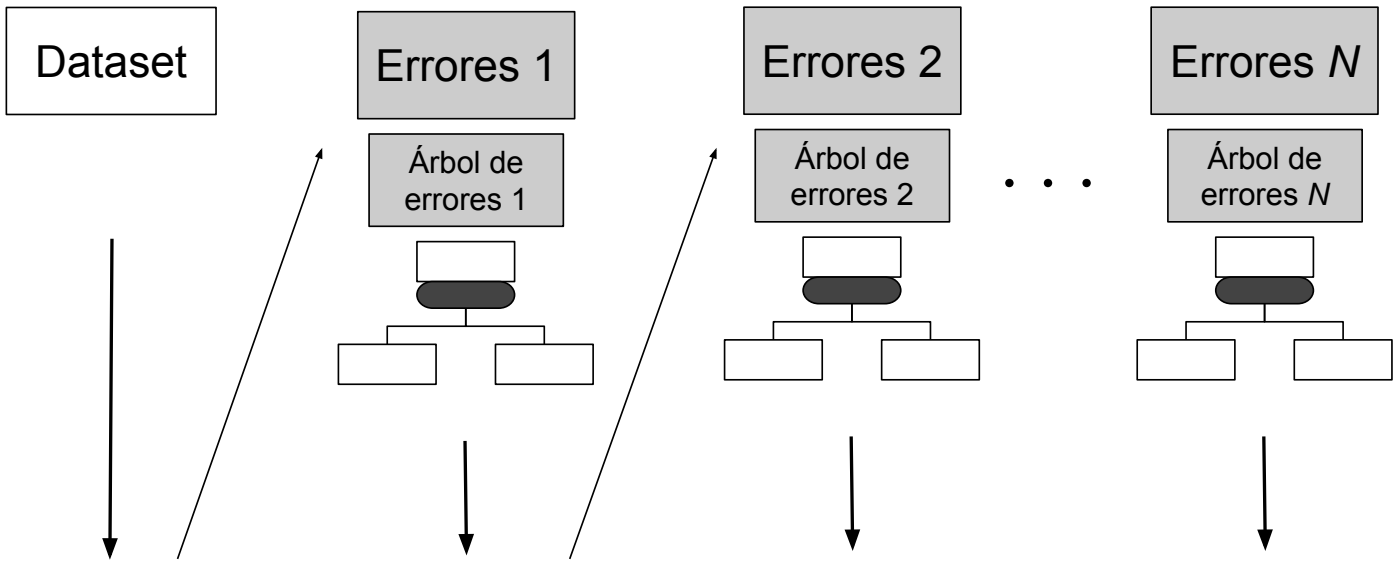
Los CART están limitados: predicciones deficientes con datos ruidosos o modelos sobreajustados.
Por ello, nacen métodos derivados más sofisticados que los potencian:



Bagging: se construyen árboles independientes, cada uno entrenado con una muestra diferente (bootstrap), y devuelve la media (regresión) o la moda (clasificación).



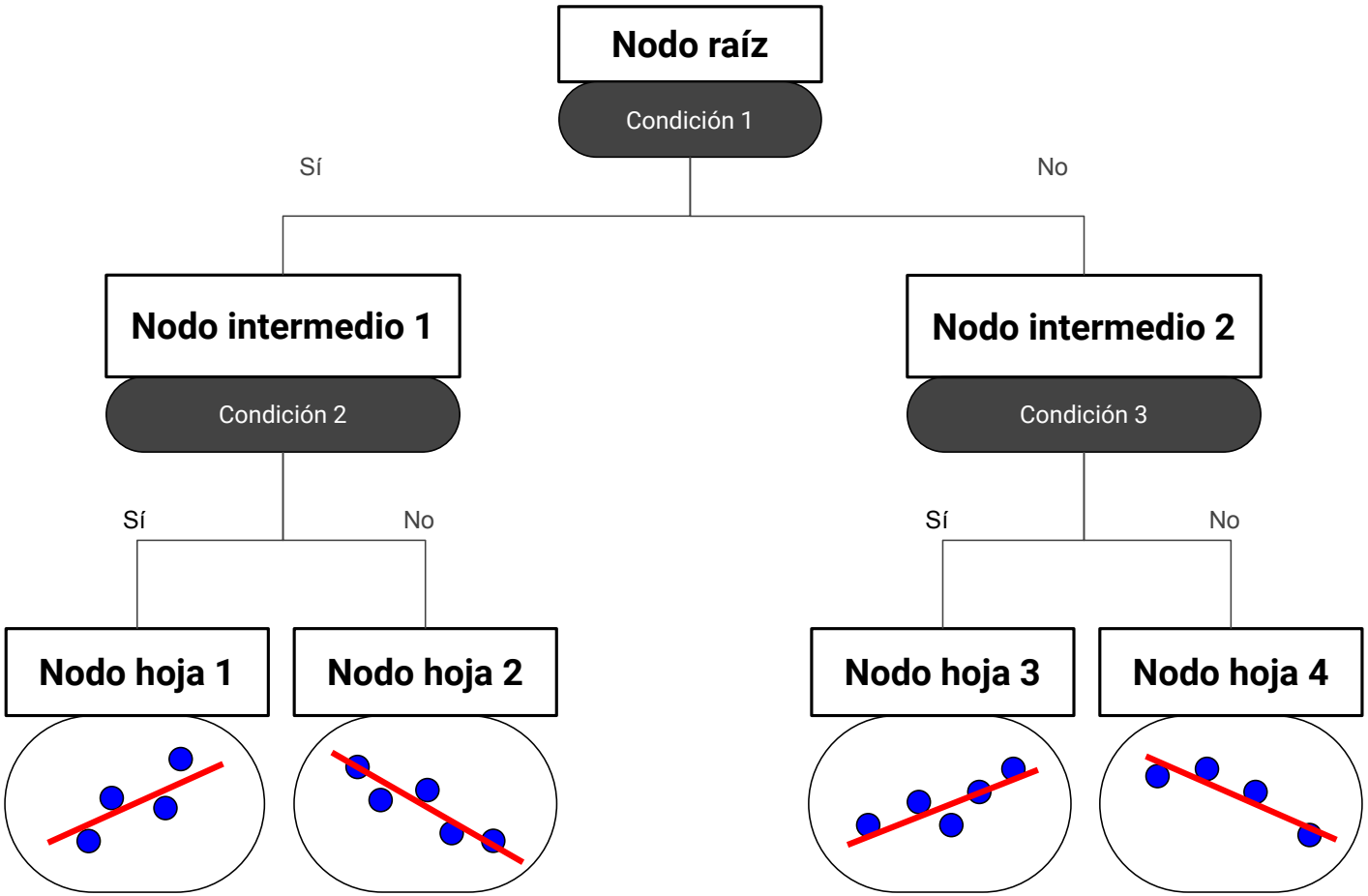
Boosting: se construye una secuencia de árboles que corrigen sus errores para realizar predicciones muy precisas.



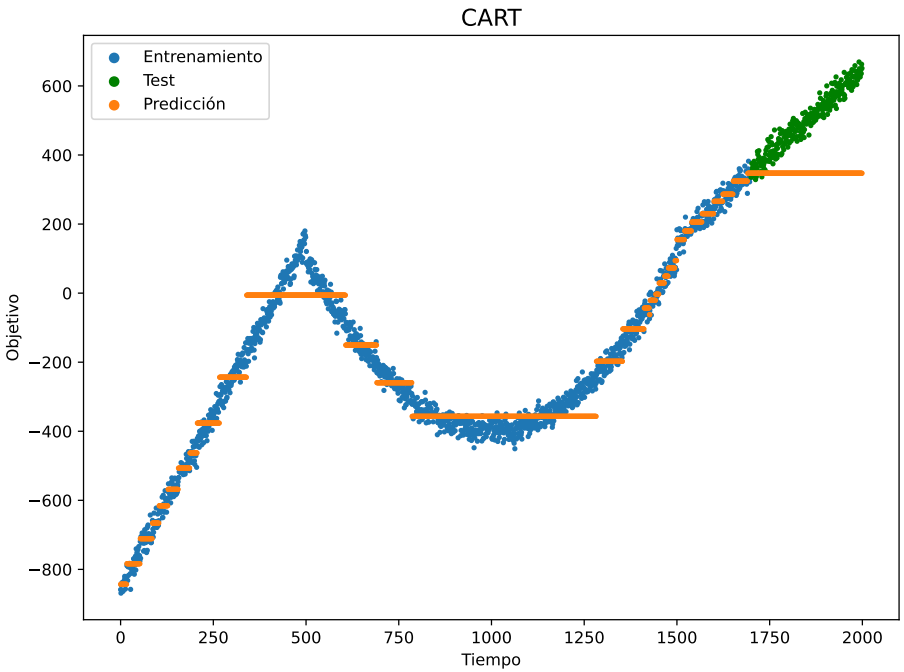
$$\hat{y} = \bar{y} + \hat{\varepsilon}_1 + \hat{\varepsilon}_2 + \dots + \hat{\varepsilon}_N$$

Algoritmos híbridos: modelos que combinan métodos basados en árboles de decisión con Regresión Lineal/Logística:

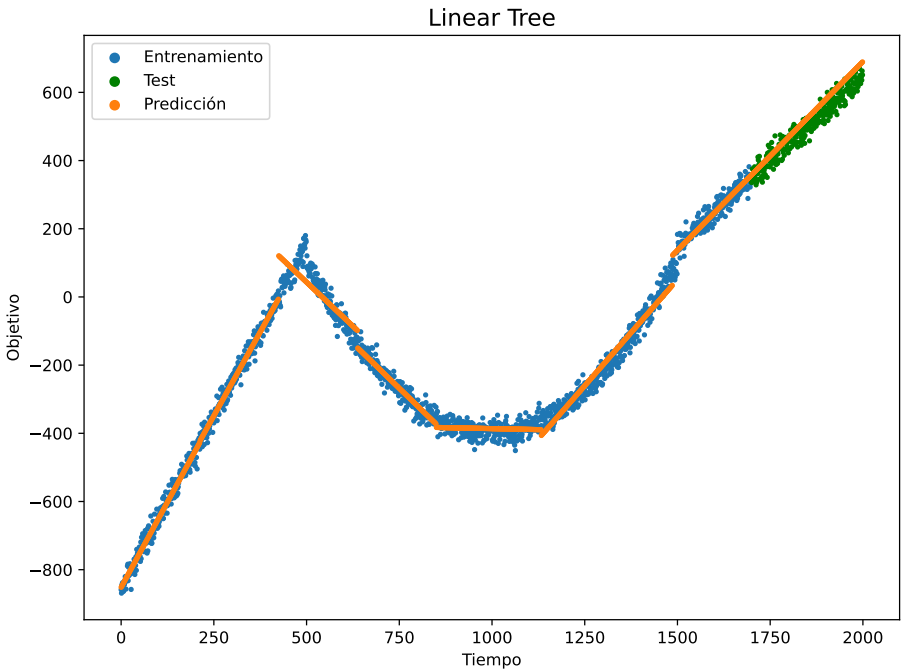
- Linear Trees
- Linear Random Forest
- Regression-Enhanced Random Forest
- Explainable Boosted Regression
- Piecewise Linear Gradient Boosting


Nodo raíz
Condición 1
Sí
No
Nodo intermedio 1
Condición 2
Nodo intermedio 2
Condición 3
Sí
No
Sí
No
Nodo hoja 1
Nodo hoja 2
Nodo hoja 3
Nodo hoja 4

En una serie temporal, el Linear Tree puede extrapolar las predicciones; el CART, no

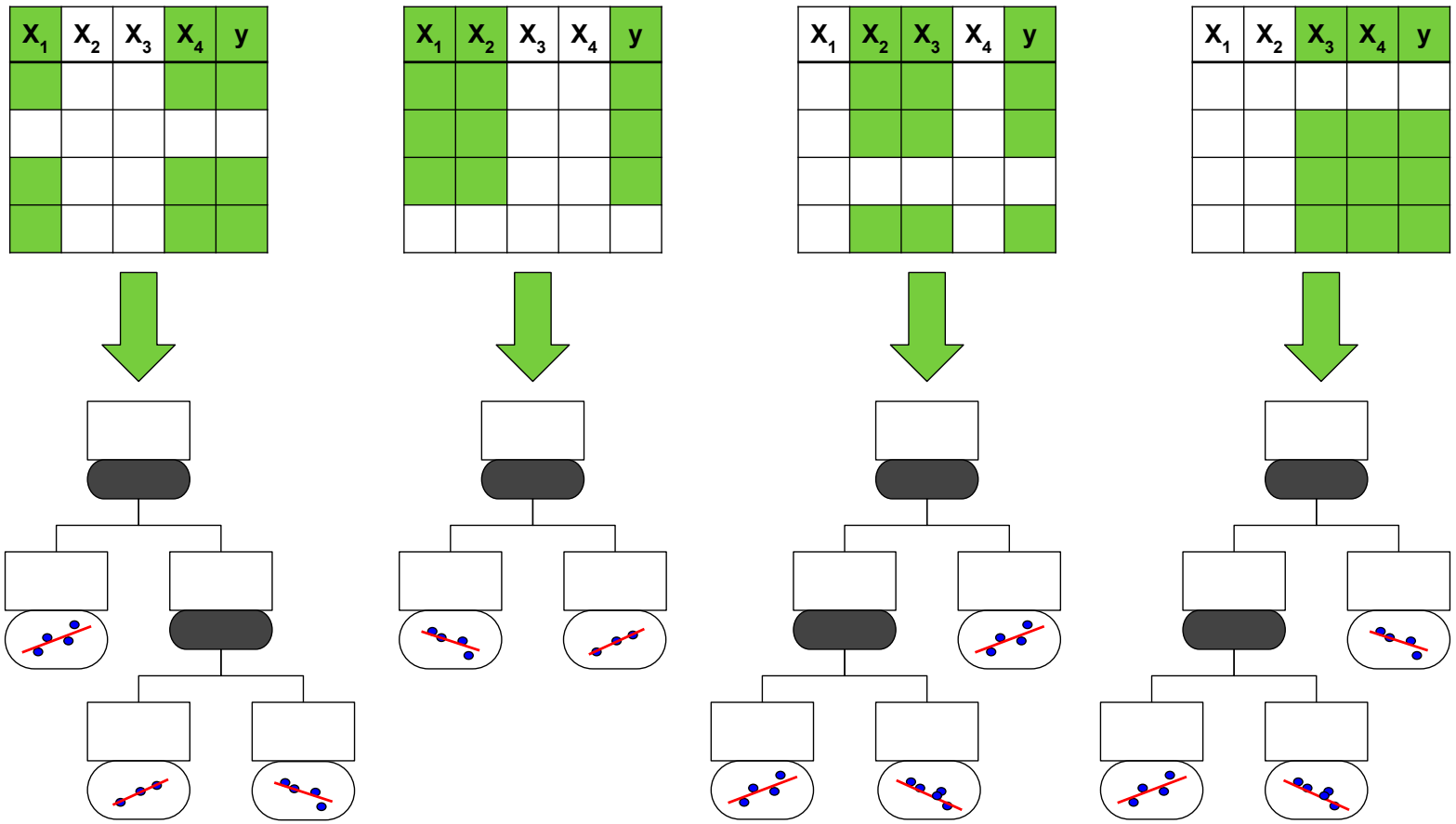


Profundidad máxima: 5



Profundidad máxima: 3

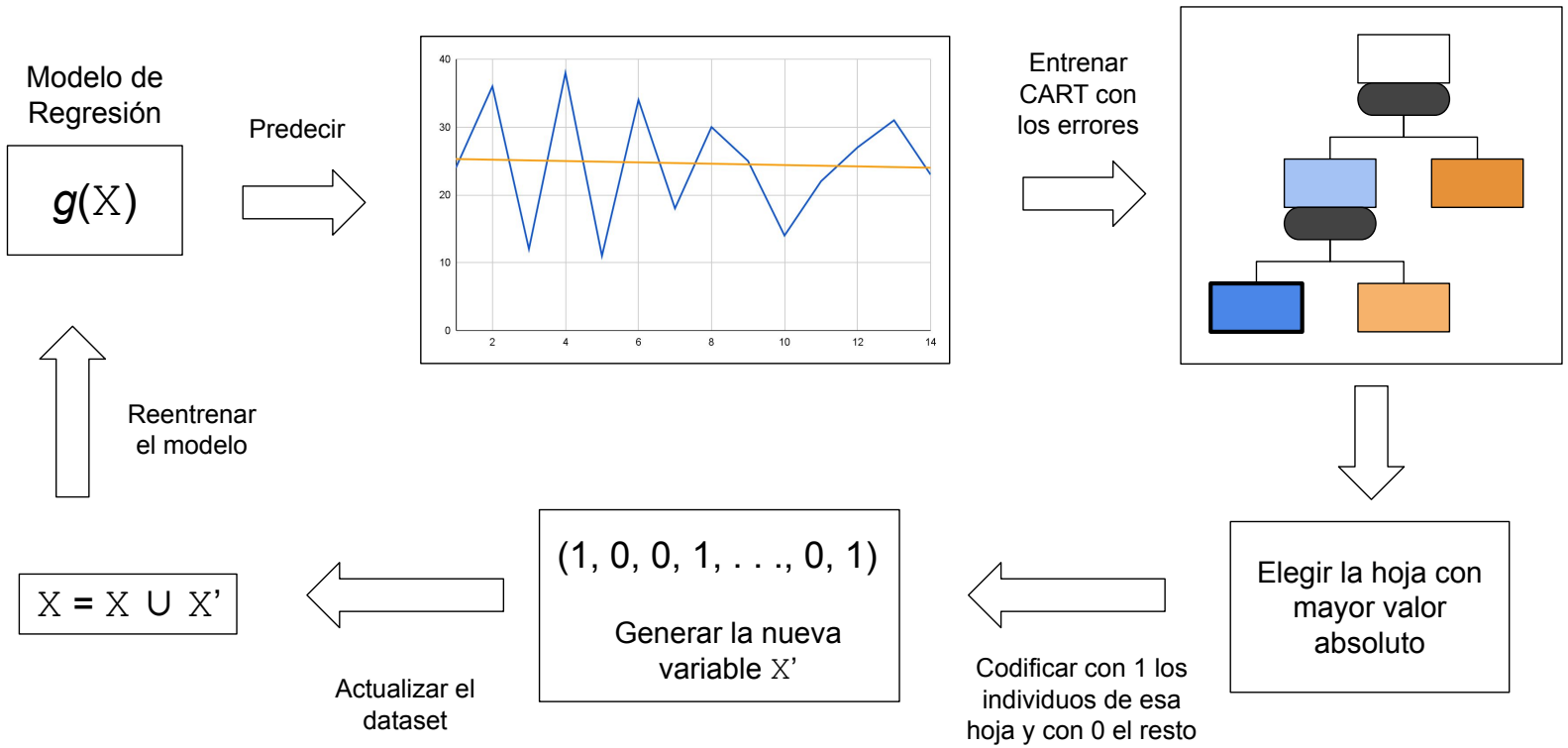
Linear Random Forest



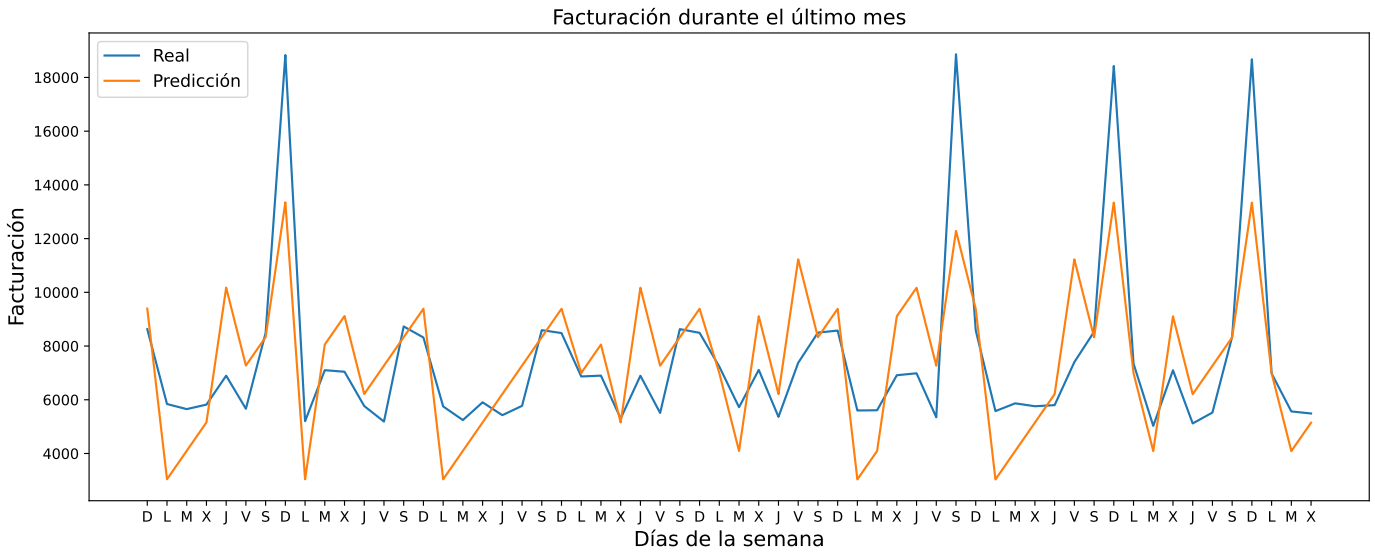
1. Se parte de las predicciones de un modelo de Regresión.
2. Se modelan sus errores con un Random Forest.
3. Se corrigen las estimaciones iniciales sumando ambas predicciones.

$$\hat{y}_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}}_{\text{Regresión Lineal/Logística}} + \underbrace{\hat{\varepsilon}_i}_{\text{Random Forest}}$$

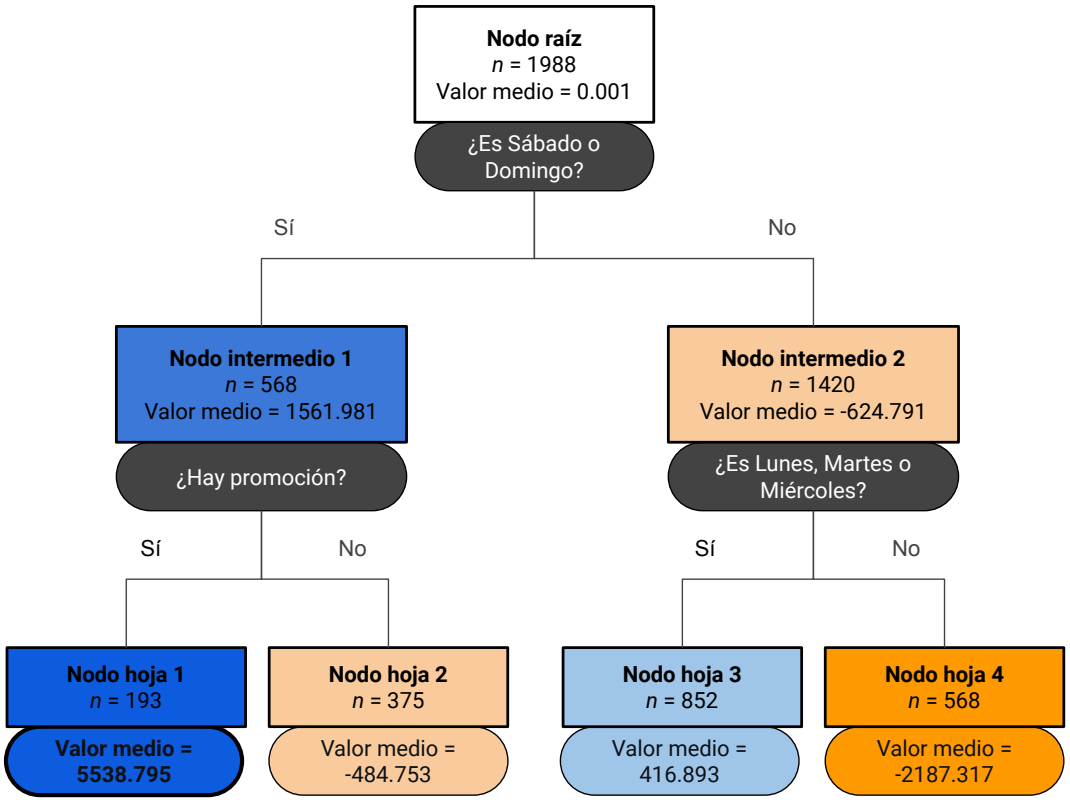
Enriquece un modelo de Regresión mediante la creación de nuevas variables dadas por un CART basadas en los errores de predicción.



Fecha	Día de la semana	Promoción	Facturación
736330	Domingo	No	8165.73
736331	Lunes	Sí	6230.86
736332	Martes	No	5226.25
⋮	⋮	⋮	⋮

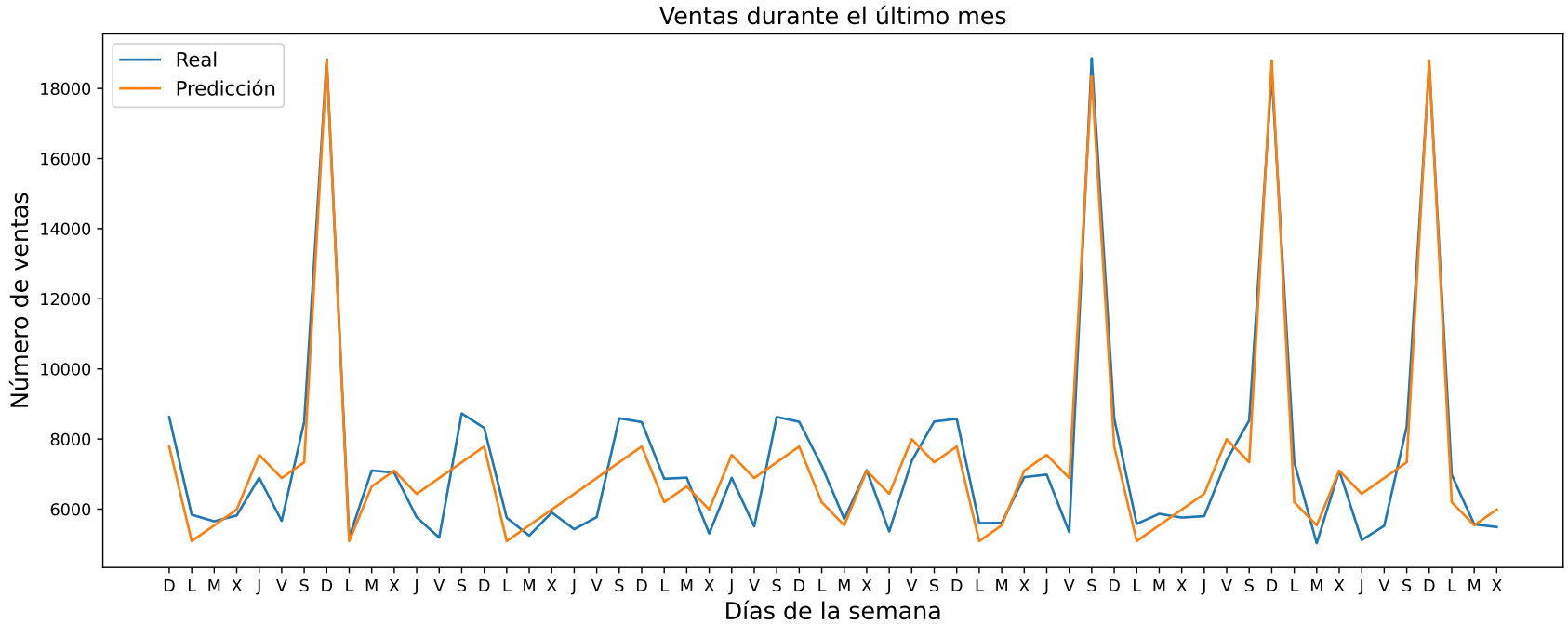


Explainable Boosted Regression

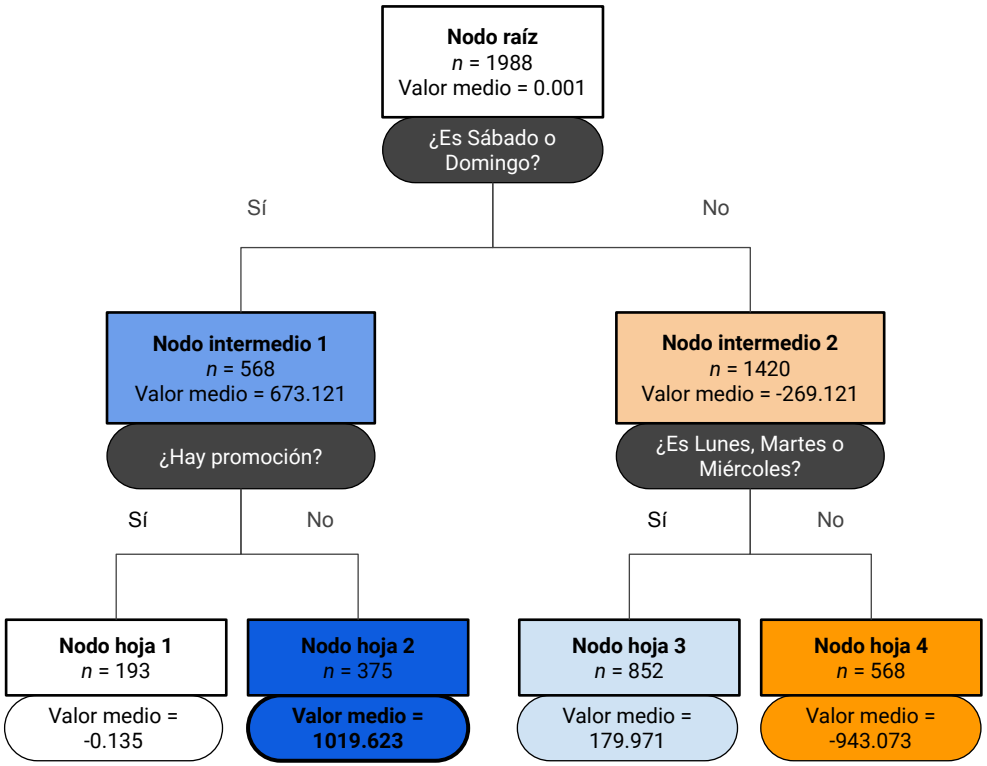


Fecha	Día de la semana	Promoción	Nueva variable 1
736330	Domingo	No	0
736331	Lunes	Sí	0
736332	Martes	No	0
⋮	⋮	⋮	⋮
738268	Sábado	Sí	1
738269	Domingo	Sí	1
⋮	⋮	⋮	⋮
738316	Viernes	No	0
738317	Sábado	No	0

Se reentrena el modelo de Regresión con la nueva variable y se obtienen las predicciones siguientes:

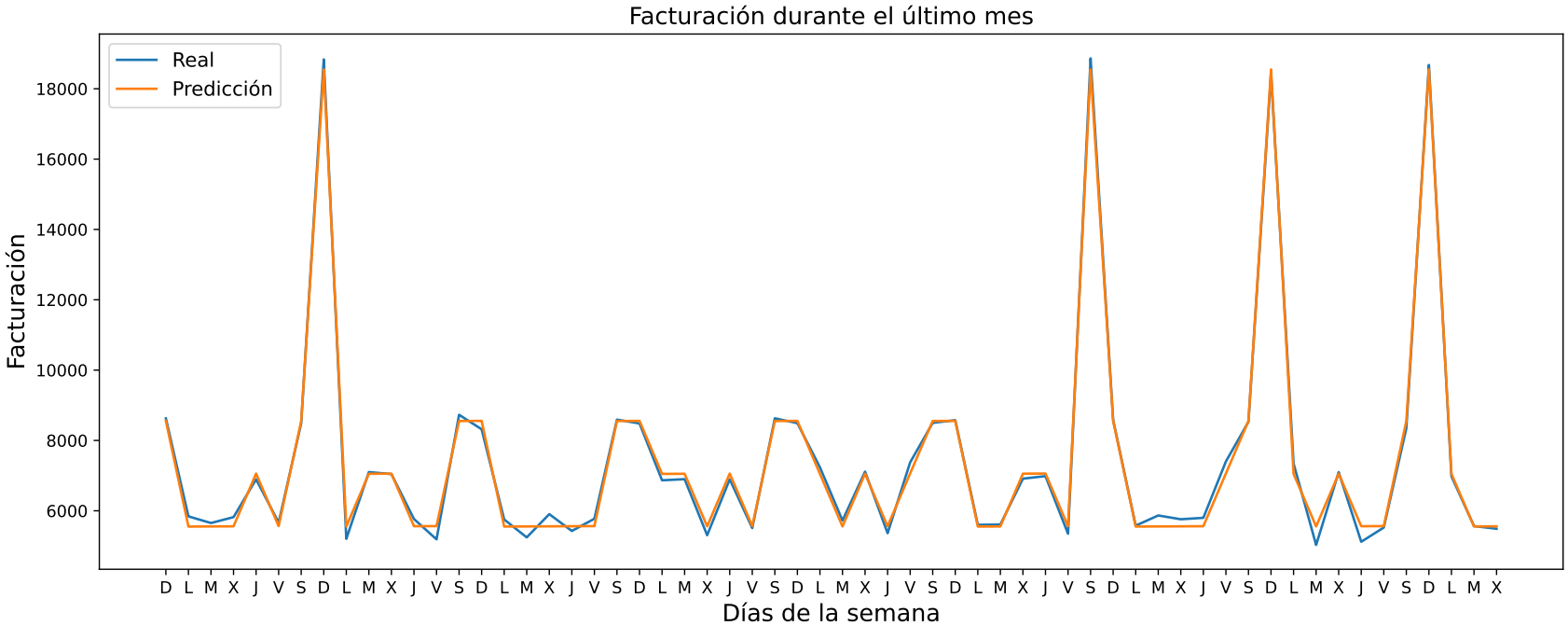


Explainable Boosted Regression

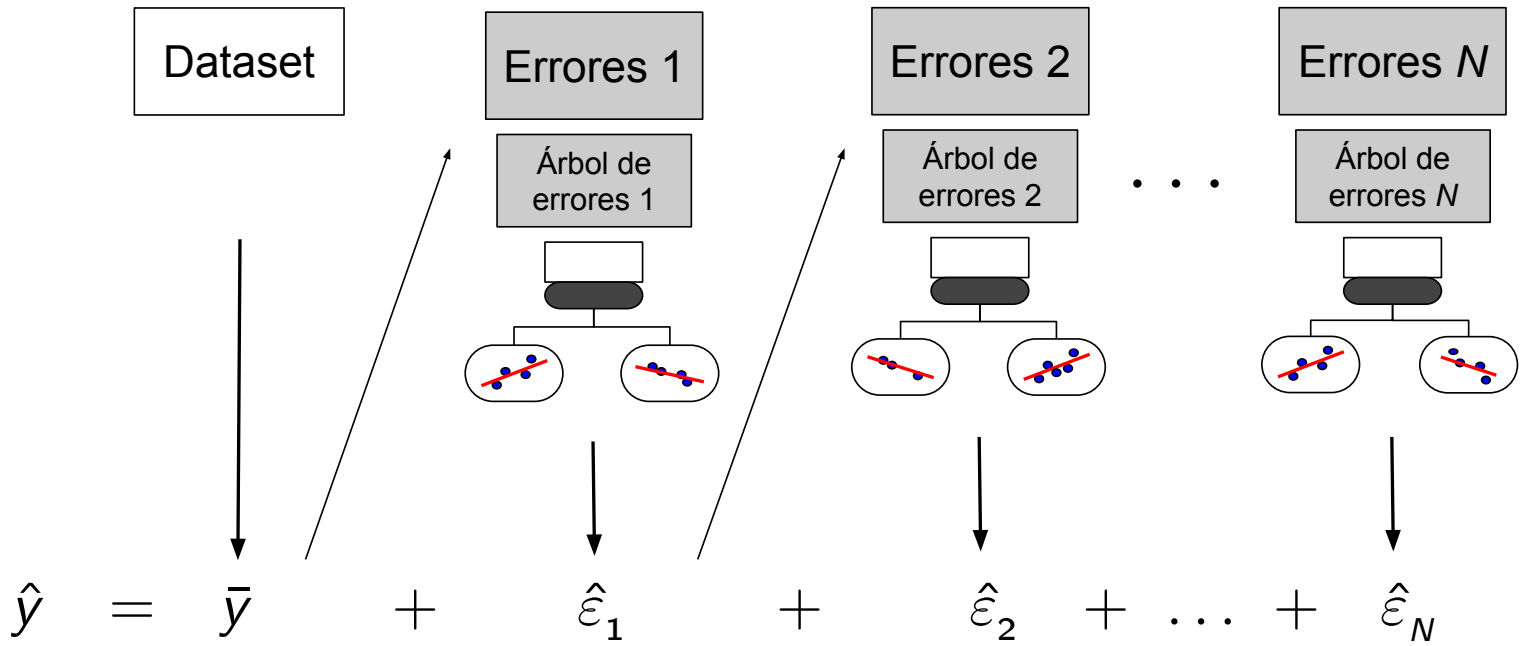


Fecha	Día de la semana	Promoción	Nueva variable 1	Nueva variable 2
736330	Domingo	No	0	1
736331	Lunes	Sí	0	0
736332	Martes	No	0	0
⋮	⋮	⋮	⋮	⋮
738268	Sábado	Sí	1	0
738269	Domingo	Sí	1	0
⋮	⋮	⋮	⋮	⋮
738316	Viernes	No	0	0
738317	Sábado	No	0	1

Se vuelve a reentrenar el modelo de Regresión con las nuevas variables y se logran las predicciones finales:



Funciona como los boosting anteriores, pero no corrigen CARTs, sino Linear Trees.



Además, emplea paralelización, ajustes de Regresión subóptimos y selección incremental de variables.

- División de los datos:
  - Conjunto de entrenamiento (70 %).
  - Conjunto de validación (15 %).
  - Conjunto de testeo (15 %).

- Métricas:

Para regresión: RMSE;

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \ .$$

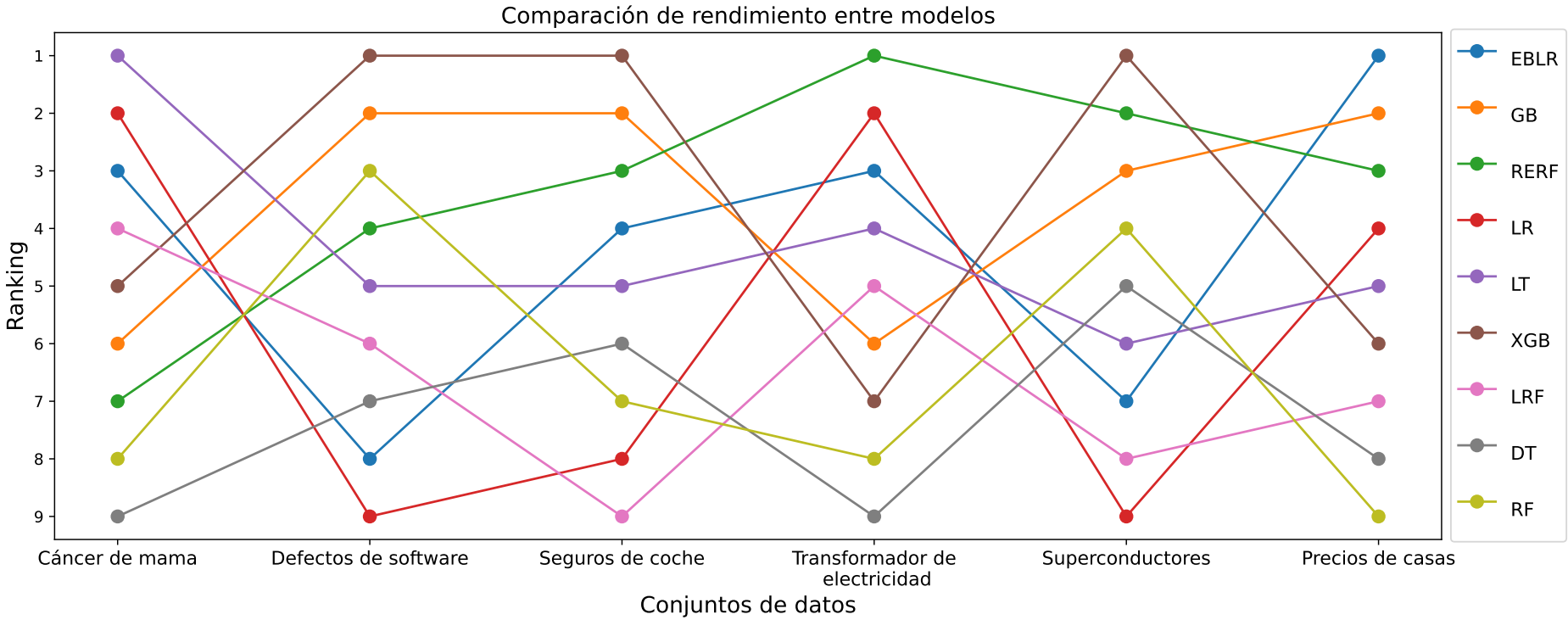
Para clasificación: AUC;

área bajo la curva ROC, que representa

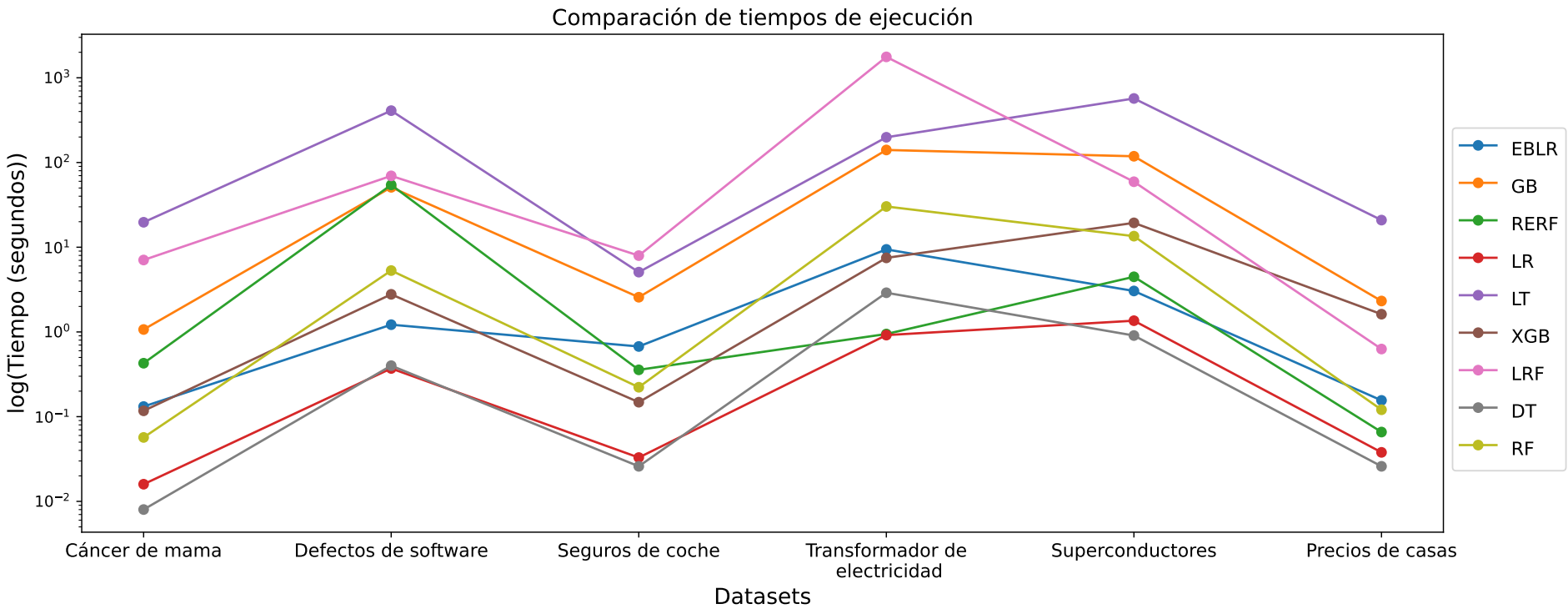
$$\text{TPR} = \frac{\text{Verdaderos Positivos}}{\text{Positivos}} \text{ frente a}$$

$$\text{TNR} = \frac{\text{Verdaderos Negativos}}{\text{Negativos}} \ .$$

Comparación en términos de precisión entre los modelos en los seis casos de uso en los datos de testeo:



Comparación entre modelos



Conclusiones

- Profundizar en Machine Learning y Ciencia de Datos.
- Descubrir los algoritmos predictivos más vanguardistas.
- Preparatorio para estudios de máster o inserción en el mundo laboral.