MODELOS DE REGRESIÓN

Manuel Valdivia

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

1. Regresión Lineal

La regresión lineal simple consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables. A la variable dependiente o respuesta se le identifica como Y, a la variable predictora o independiente como X.

El modelo de regresión lineal simple se describe de acuerdo a la ecuación:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

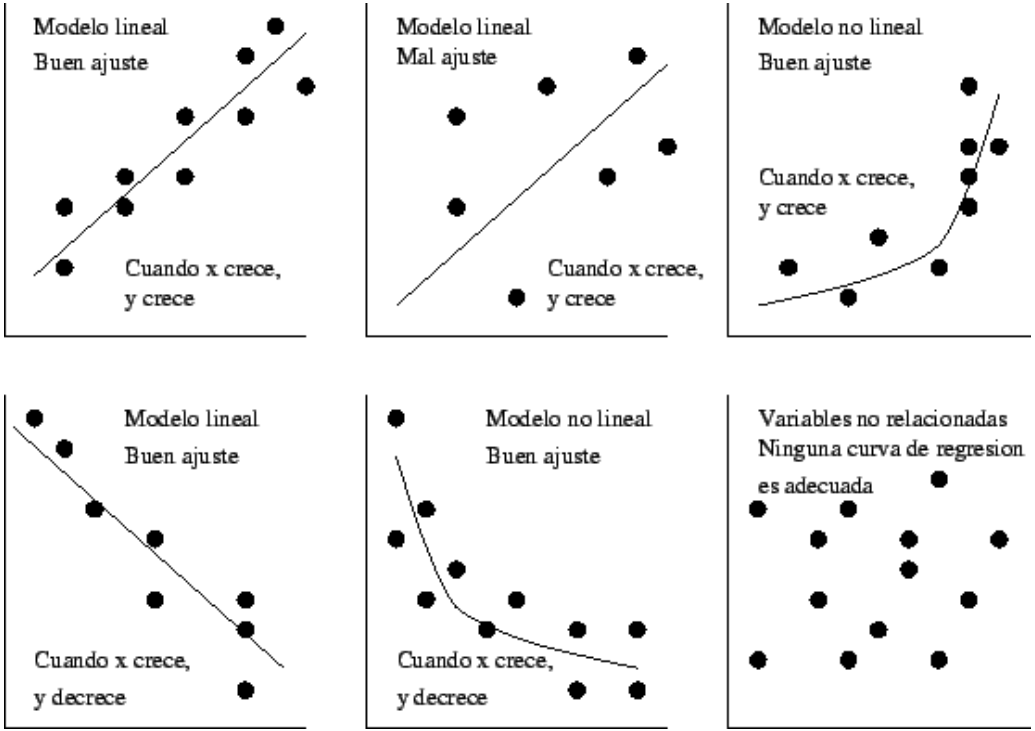
Siendo:
- la ordenada en el origen β_0
- la pendiente β_1
- el error aleatorio. ϵ

Este último representa la diferencia entre el valor ajustado por la recta y el valor real.

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

1. Regresión Lineal



¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

1. Regresión Lineal

Condiciones:

- Linealidad: La relación entre ambas variables debe ser lineal.
- Distribución Normal de los residuos: Los residuos se tiene que distribuir de forma normal, con media igual a 0.
- Varianza de residuos constante (homocedasticidad): La varianza de los residuos ha de ser aproximadamente constante a lo largo del eje X.
- Valores atípicos y de alta influencia: Hay que estudiar con detenimiento los valores atípicos o extremos ya que pueden generar una falsa correlación que realmente no existe, u ocultar una existente.
- Independencia, Autocorrelación: Las observaciones deben ser independientes unas de otras.

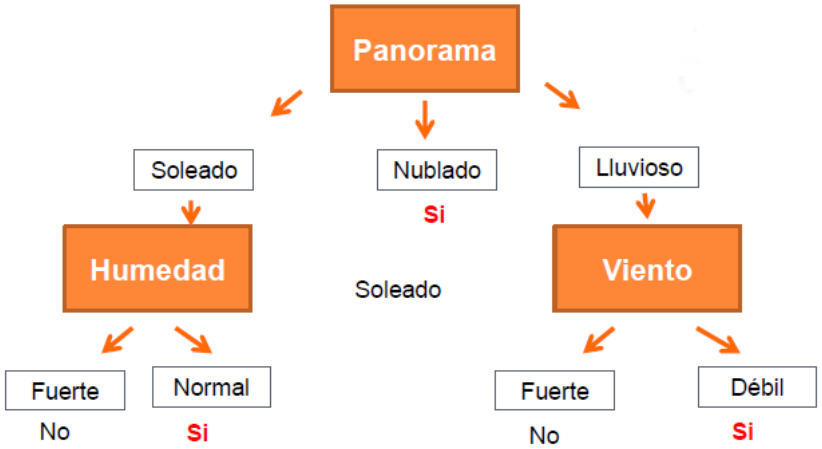
¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

2. Arboles de Clasificación, Arboles de Regresión

Un Árbol de decisión es un modelo de predicción que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva

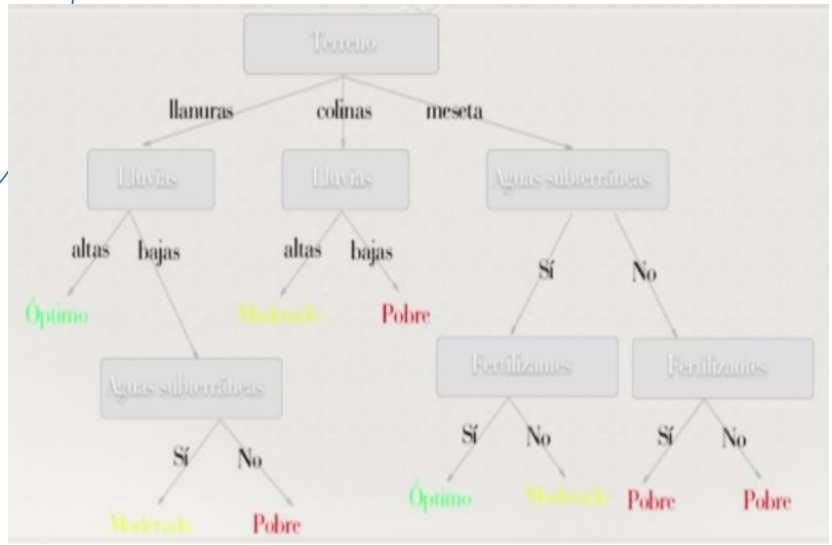
Ejemplo: la pregunta a responder es Juego Tennis?



¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

2. Arboles de Clasificación, Arboles de Regresión



- ***Cosecha = Óptima***
  - ***Terreno = Llanura y Lluvias = Altas***
  - ***0 Terreno = Meseta y Aguas Subterráneas = Sí y Fertilizantes = Sí***
- ***Cosecha = Moderada***
  - ***Terreno = Llanura y Lluvias = Bajas y Aguas Subterráneas= Sí***
  - ***0 Terreno = Colinas y Lluvias = Altas***
- ***0 Terreno = Meseta y Aguas Subterráneas = Sí y Fertilizantes = No***

Llúvias	Óptima	Moderada	Pobre
Altas	3 (43%)	4 (57%)	0
Medias	1 (17%)	2 (33%)	3 (50%)
Bajas	0	2 (29%)	5 (71%)

Terreno	Óptima	Moderada	Pobre
Colinas	0	3 (50%)	3 (50%)
Llanuras	2 (33%)	1 (17%)	3 (50%)
Mesetas	2 (25%)	3 (37.5%)	3 (37.5%)

Fertilizante	Óptima	Moderada	Pobre
No	0	6 (67%)	3 (33%)
Sí	4 (36%)	1 (9%)	6 (55%)

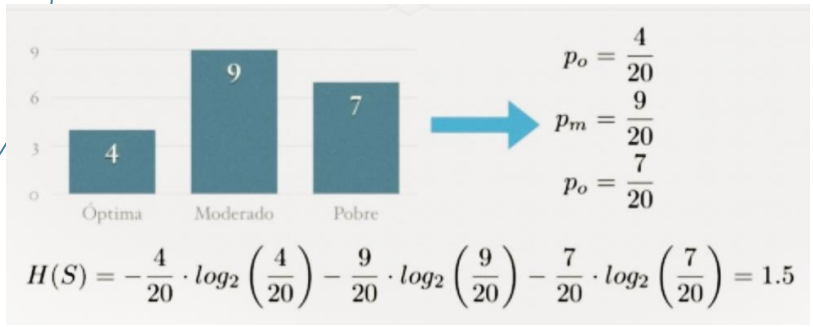
Agua Sub.	Óptima	Moderada	Pobre
No	2 (18%)	6 (55%)	3 (27%)
Si	2 (22%)	1 (11%)	6 (67%)

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

2. Arboles de Clasificación, Arboles de Regresión

Entropía



$$p_o = \frac{4}{20}$$
$$p_m = \frac{9}{20}$$
$$p_o = \frac{7}{20}$$

$$H(S) = -\frac{4}{20} \cdot log_2\left(\frac{4}{20}\right) - \frac{9}{20} \cdot log_2\left(\frac{9}{20}\right) - \frac{7}{20} \cdot log_2\left(\frac{7}{20}\right) = 1.5$$

Ganancia de Información

$$\Delta H(S,V) = H(S) - \sum_{c \in V} \frac{|V = c|}{|V|} H(V = c)$$

$Terreno = \{llanura, meseta, colina\}$

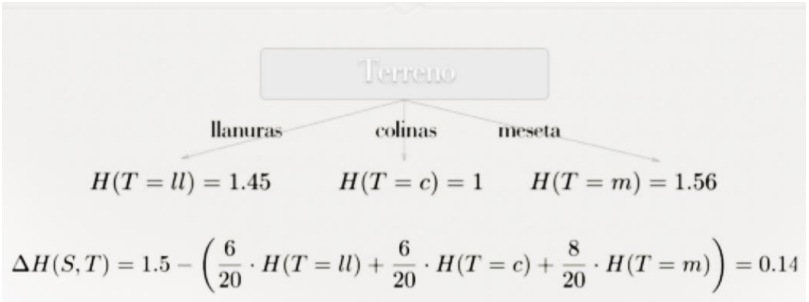
Terreno	Óptima	Moderada	Pobre
Colinas	0	3 (50%)	3 (50%)
Llanuras	2 (33%)	1 (17%)	3 (50%)
Mesetas	2 (25%)	3 (37.5%)	3 (37.5%)

$$H(T = c) = -\left(\frac{0}{6}log_2\left(\frac{0}{6}\right) + \frac{3}{6}log_2\left(\frac{3}{6}\right) + \frac{3}{6}log_2\left(\frac{3}{6}\right)\right) = 1$$

$$H(T = ll) = -\left(\frac{2}{6}log_2\left(\frac{2}{6}\right) + \frac{1}{6}log_2\left(\frac{1}{6}\right) + \frac{3}{6}log_2\left(\frac{3}{6}\right)\right) = 1.45$$

$$H(T = m) = -\left(\frac{2}{8}log_2\left(\frac{2}{8}\right) + \frac{3}{8}log_2\left(\frac{3}{8}\right) + \frac{3}{8}log_2\left(\frac{3}{8}\right)\right) = 1.56$$

Ganancia de Información



$$H(T = ll) = 1.45 \qquad H(T = c) = 1 \qquad H(T = m) = 1.56$$

$$\Delta H(S,T) = 1.5 - \left(\frac{6}{20} \cdot H(T = ll) + \frac{6}{20} \cdot H(T = c) + \frac{8}{20} \cdot H(T = m)\right) = 0.14$$

Ganancia de Información

Parámetro	Ganancia de Información
Terreno	0.14
Lluvia	0.42
Fertilizantes	0.36
Aguas Subterráneas	0.16

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

2. Arboles de Clasificación, Arboles de Regresión

Índice de Gini

Para cada variable tenemos

Categoría	C1	C2
Sí	2	6
No	8	4
Total	10	10

$$Gini_{C_1} = (0.2)^2 + (0.8)^2 = 0.68$$
$$Gini_{C_2} = (0.6)^2 + (0.4)^2 = 0.52$$
$$GiniIndex = \frac{10}{20} \cdot 0.68 + \frac{10}{20} \cdot 0.52 = 0.60$$

Reducción de la Varianza

Para cada variable tenemos

Categoría	C1	C2	Total
1	2	6	8
0	8	4	12
Total	10	10	20

Convertimos

- *Sí = 1*
- *No = 0*

Reducción de la Varianza

Para cada variable llevamos a cabo

Categoría	C1	C2	Total
1	2	6	8
0	8	4	12
Total	10	10	20

$$mean(root) = \frac{8 \cdot 1 + 12 \cdot 0}{20} = 0.4$$
$$var(root) = \frac{8 \cdot (1 - 0.4)^2 + 12 \cdot (0 - 0.4)^2}{20} = 0.24$$
$$mean(C_1) = \frac{2 \cdot 1 + 8 \cdot 0}{10} = 0.2$$
$$Var(C_1) = \frac{2 \cdot (1 - 0.2)^2 + 8 \cdot (0 - 0.2)^2}{10} = 0.16$$
$$mean(C_2) = \frac{6 \cdot 1 + 4 \cdot 0}{10} = 0.6$$
$$Var(C_2) = \frac{6 \cdot (1 - 0.6)^2 + 4 \cdot (0 - 0.6)^2}{10} = 0.24$$
$$Var.Ponderada = \frac{10}{20} \cdot 0.16 + \frac{10}{20} \cdot 0.24 = 0.20$$

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

2. Arboles de Clasificación, Arboles de Regresión

Algoritmos:

- CART
- CHAID
- ID3
- C50

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

Algoritmo CART

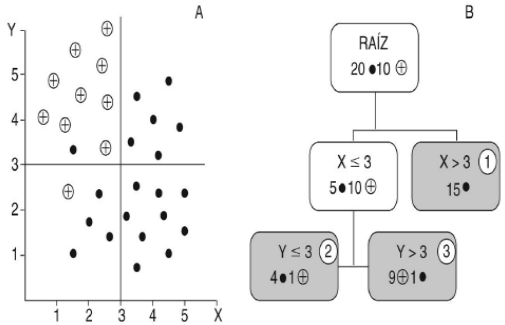
Notas:

- Árboles de clasificación: predicen categorías de objetos.
- Árboles de regresión: predicen valores continuos.
- Partición binaria recursiva.
- En cada iteración se selecciona la variable predictiva y el punto de separación que mejor reduzcan la 'impureza'.

Características:

- Uno de los métodos de aprendizaje supervisado no paramétrico más utilizado.
- Realizar sucesivas divisiones binarias en un conjunto de datos guiado por un criterio.
- Para cada nodo selecciona a la variable independiente que proporciona el mejor desempeño en el criterio para particionar los datos.

Idea Intuitiva:



¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

Algoritmo CART

Criterios de Partición:

- Objetivo de una partición: Incrementar la homogeneidad (en términos de clase) de los subconjuntos resultantes. Que sean más puros que el conjunto originario.
- Que sean más puros que el conjunto originario. Existen criterios de impureza tales como : Medida de Entropia, Indice de Gini.

Índice de Gini:

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) G(C|A_{ij})$$

$$G(C|A_{ij}) = -\sum_{k=1}^{J} p(C_k|A_{ij}) p(\neg C_k|A_{ij}) =$$

$$= 1 - \sum_{k=1}^{J} p^2(C_k|A_{ij})$$

Criterio de Parada:

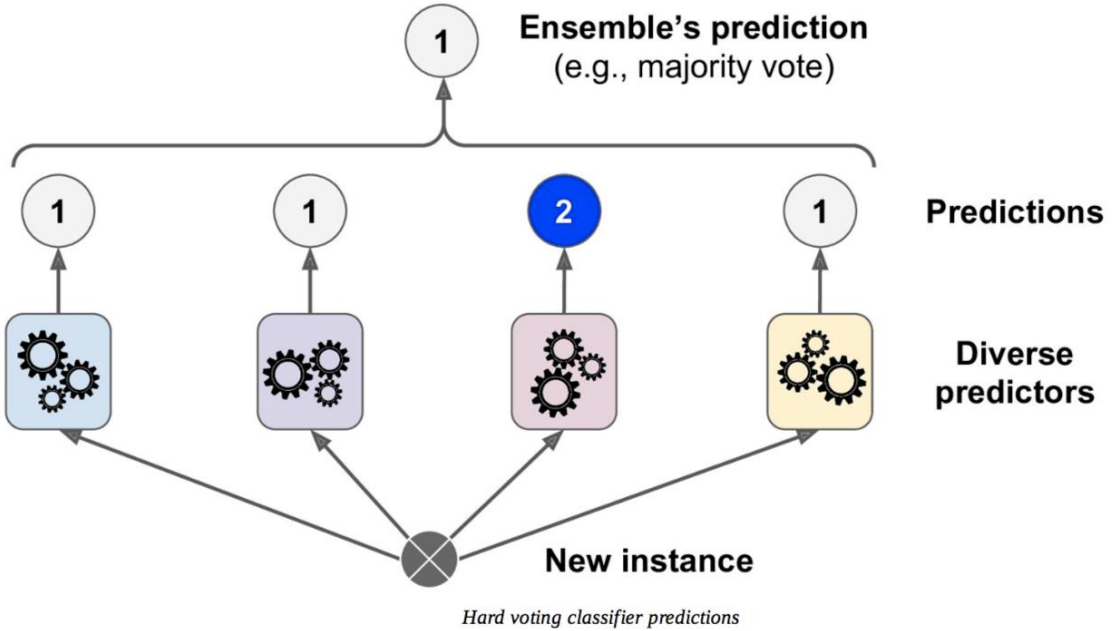
- Un nodo se declarará terminal si el nodo es puro.
- Un nodo se declarará terminal si cualquier otra subdivisión no da una mejora mayor que la obtenida en el nodo padre.

- Ai es el atributo para ramificar el árbol.
- Mi es el número de valores diferentes del atributo Ai.
- p(Aij) es la probabilidad de que Ai tome su j-ésimo valor (1 <= j <= Mi).
- p(Ck | Aij) es la probabilidad de que un ejemplo pertenezca a la clase Ck cuando su atributo Ai toma su j-ésimo valor.

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

3. Ensemble Voting



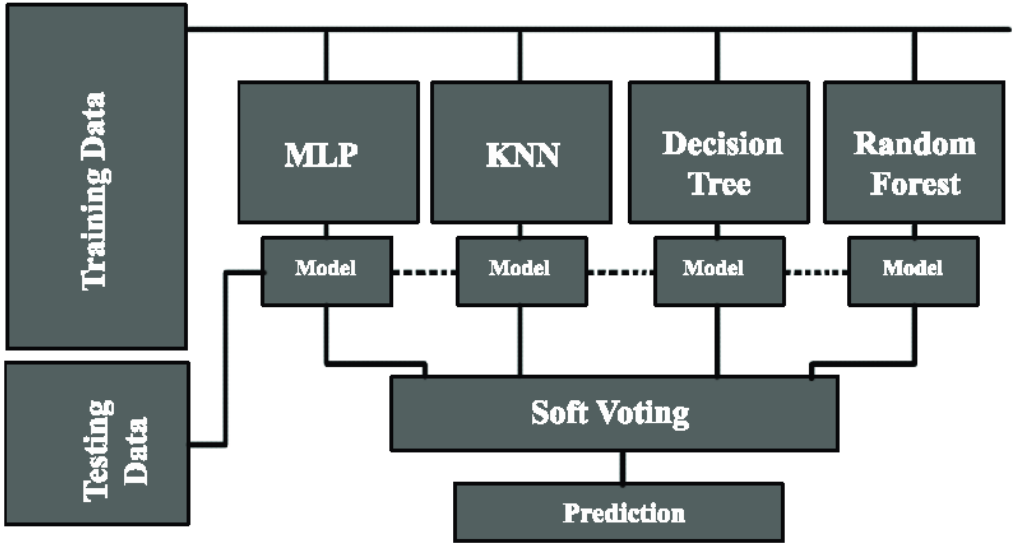
Hard voting classifier predictions

Es una técnica que se puede utilizar para mejorar el rendimiento del modelo, idealmente logrando un mejor rendimiento que cualquier modelo único utilizado en el conjunto.

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

3. Ensemble Voting

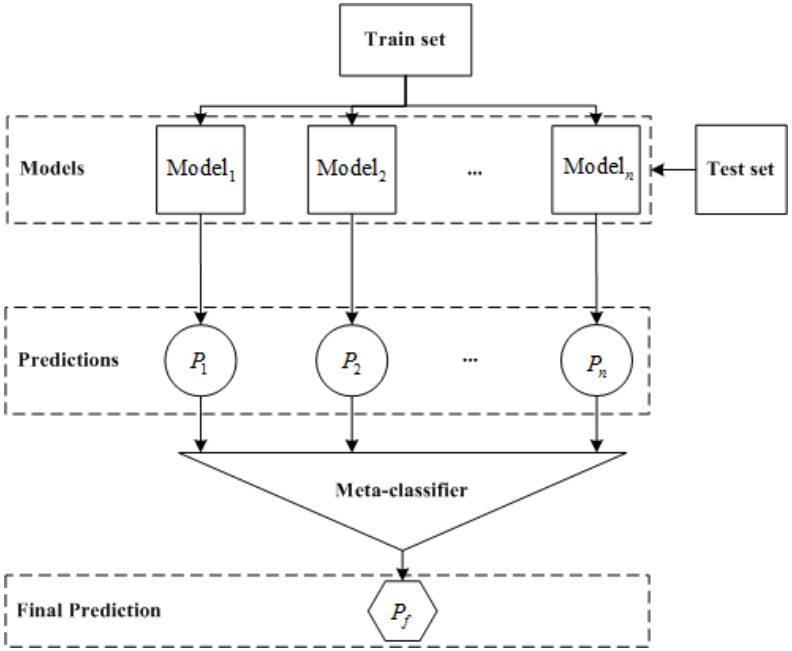


En el caso de la regresión, esto implica calcular el promedio de las predicciones de los modelos. En el caso de la clasificación, se suman las predicciones para cada etiqueta y se predice la etiqueta con el voto mayoritario.

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

4. Ensemble Stacking

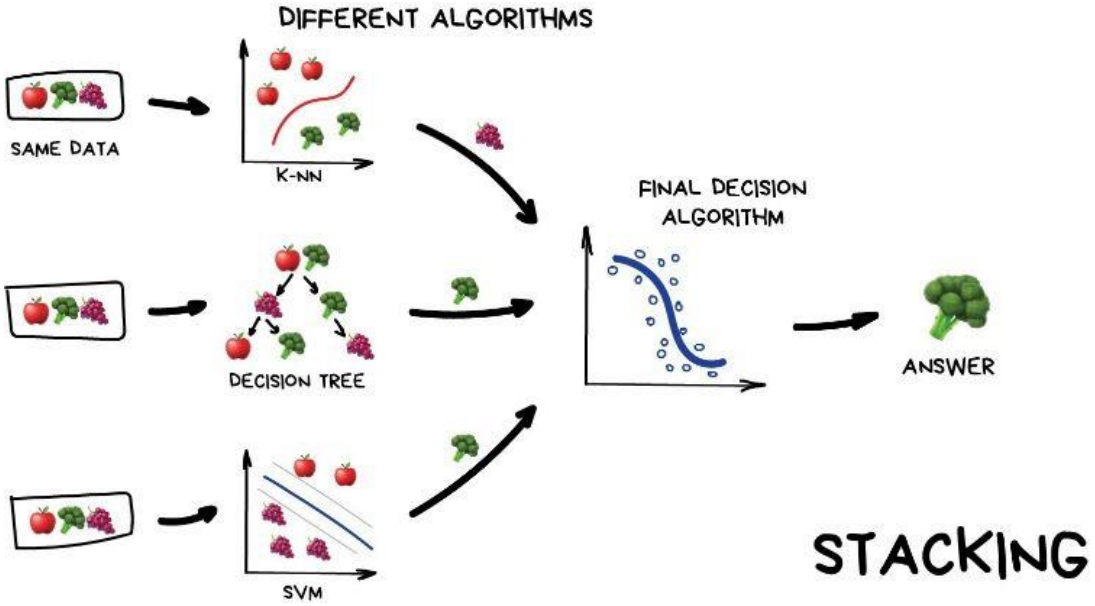


Utiliza un algoritmo de metaaprendizaje para aprender cómo combinar mejor las predicciones de dos o más algoritmos básicos de aprendizaje automático.

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

4. Ensemble Stacking

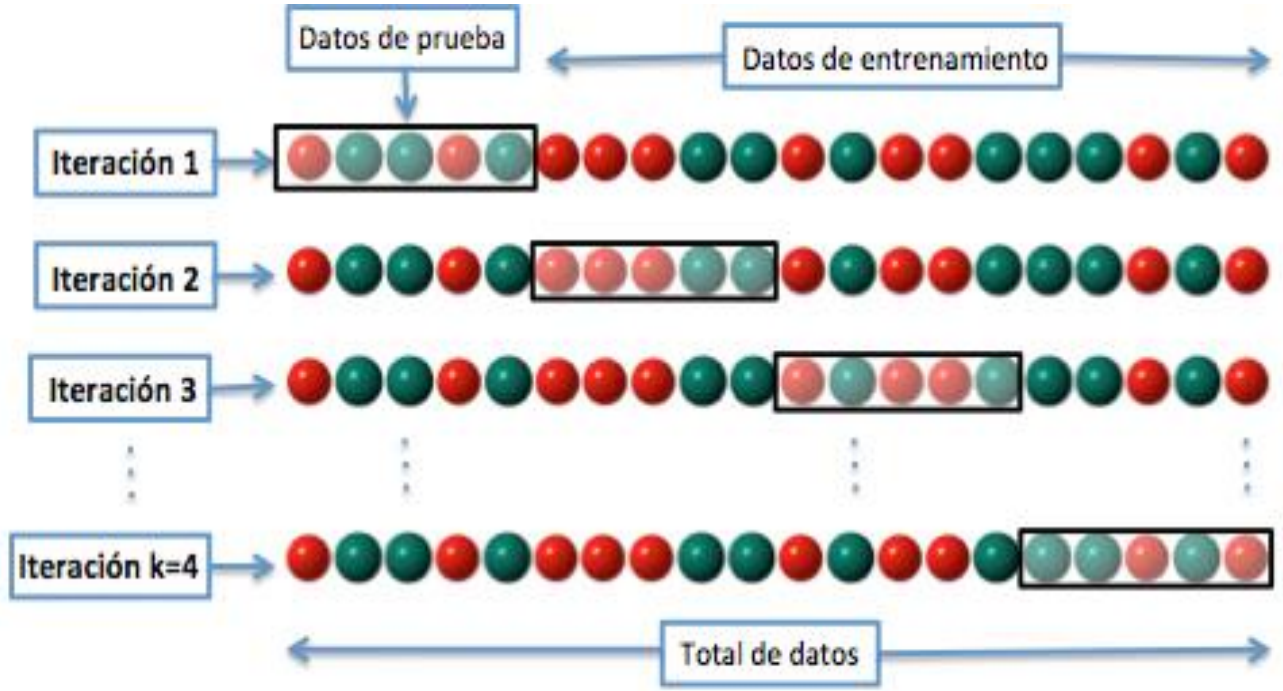


El beneficio del apilamiento es que puede aprovechar las capacidades de una variedad de modelos con buen desempeño en una tarea de clasificación o regresión y hacer predicciones que tienen un mejor desempeño que cualquier modelo individual en el conjunto.

¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

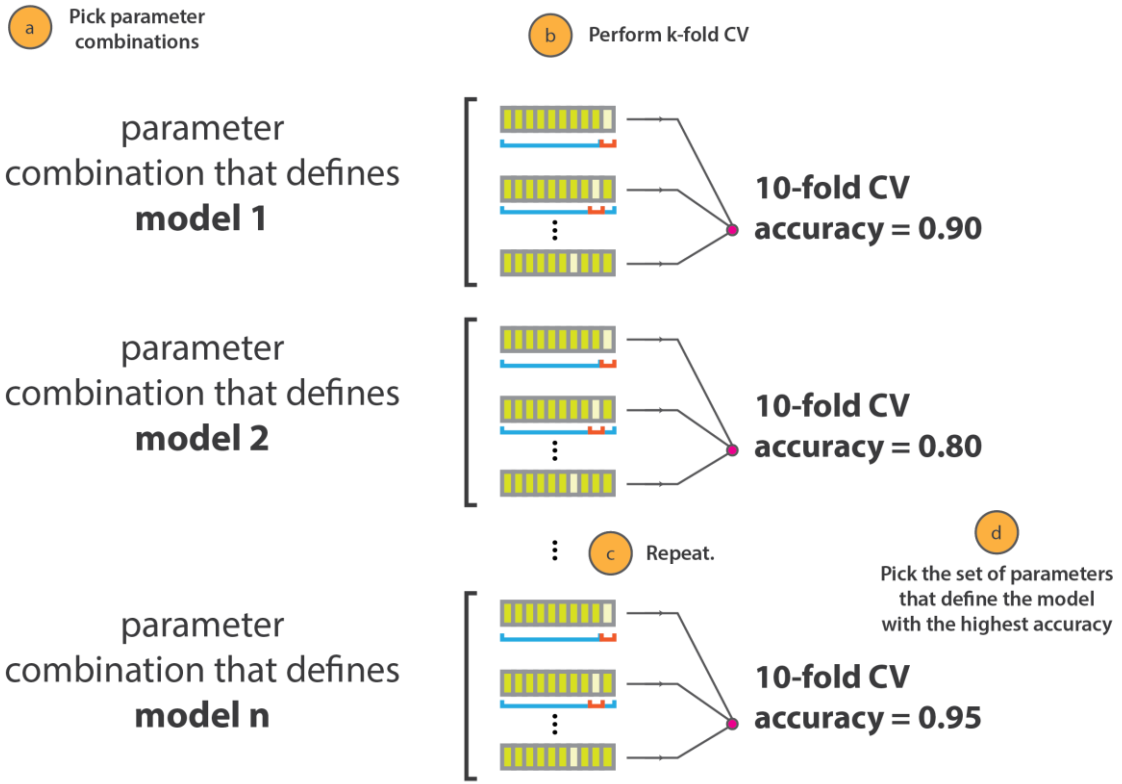
5. Cross Validation



¿Qué solución planteamos?

Data Analytics en Banca – Modelamiento de los datos

5. Cross Validation



GRACIAS