Técnicas de Modelación Predictiva o Clasificación Supervisada

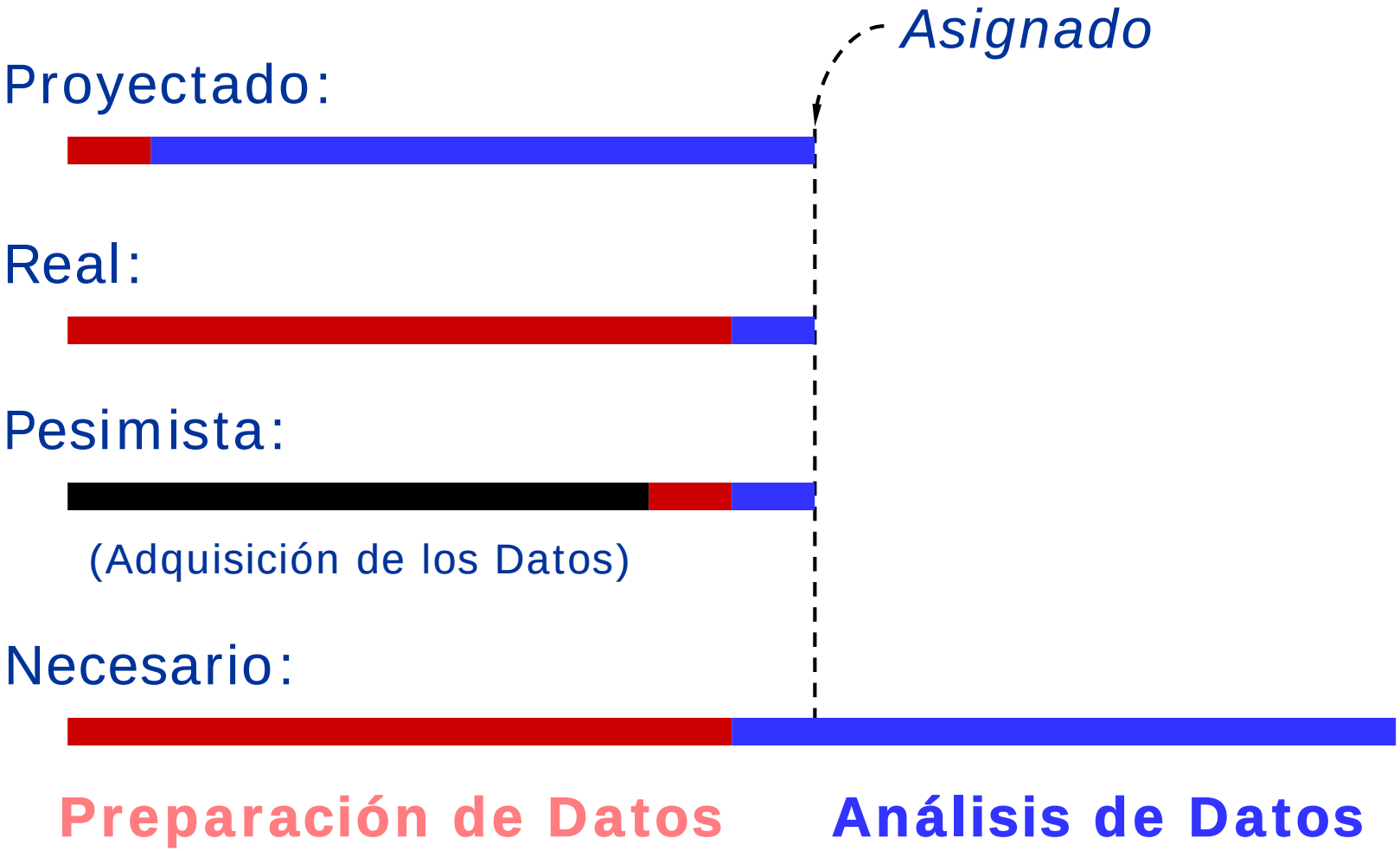
Dr. Viterbo H. Berberena González

Director de Minería de Datos

Pearson S.A. de C.V.

La Modelación y las Dificultades con los Datos

Estimación de los Tiempos



La Disposición de los Datos

Largo-estrecho

Acct	type
2133	MTG
2133	SVG
2133	CK
2653	CK
2653	SVG
3544	MTG
3544	CK
3544	MMF
3544	CD
3544	LOC

Corto-ancho

Acct	CK	SVG	MMF	CD	LOC	MTG
2133	1	1	0	0	0	1
2653	1	1	0	0	0	0
3544	1	0	1	1	1	1

La Creación de Variables de Entrada

Claim Date	Accident Time	Delay	Season	Dark
11nov96	102396/12:38	19	fall	0
22dec95	012395/01:42	333	winter	1
26apr95	042395/03:05	3	spring	1
02jul94	070294/06:25	0	summer	0
08mar96	123095/18:33	69	winter	0
15dec96	061296/18:12	186	summer	0
09nov94	110594/22:14	4	fall	1

Agrupamiento Horizontal

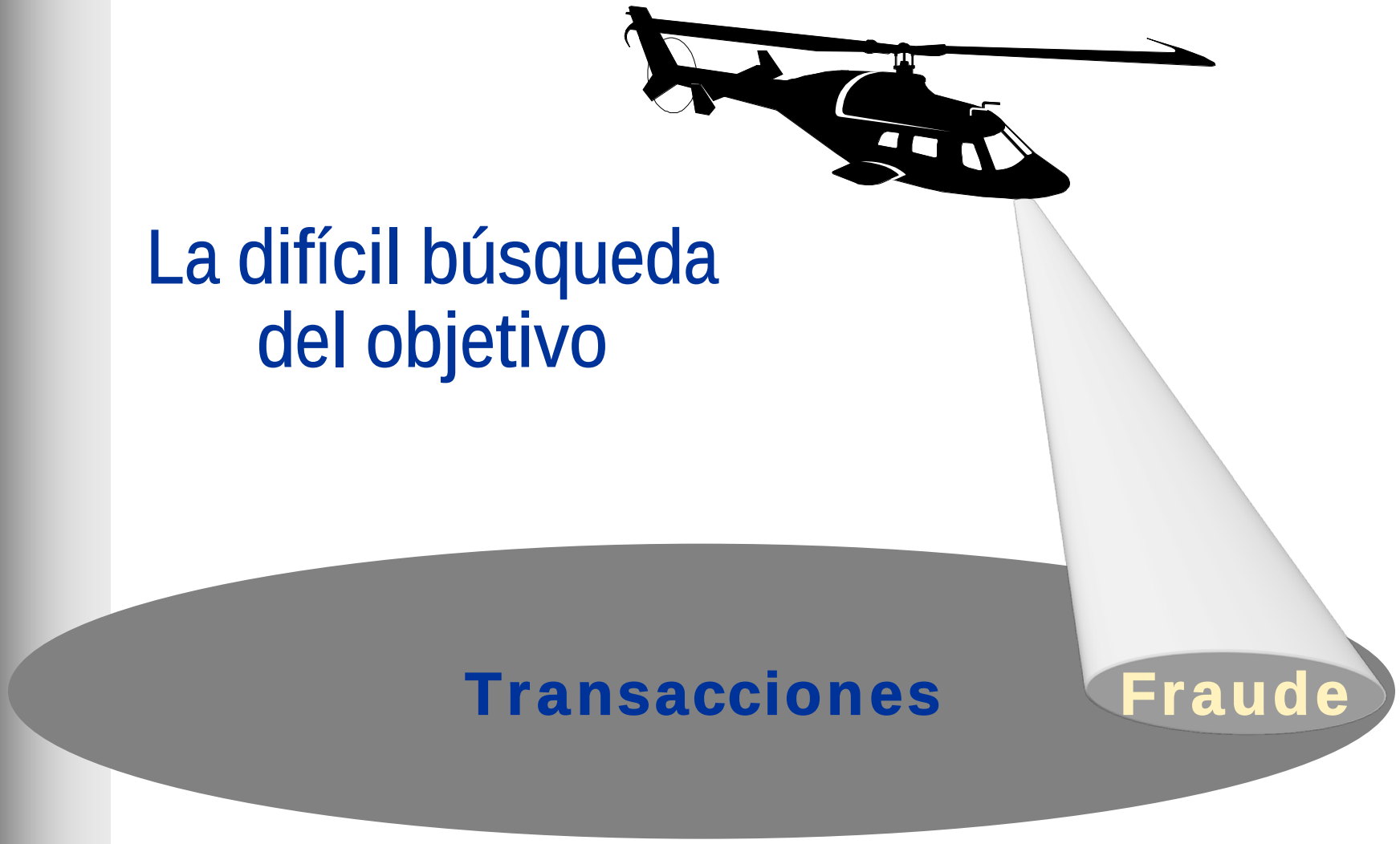
HH	Acct	Sales
4461	2133	160
4461	2244	42
4461	2773	212
4461	2653	250
4461	2801	122
4911	3544	786
5630	2496	458
5630	2635	328
6225	4244	27
6225	4165	759

HH	Acct	Sales
4461	2133	?
4911	3544	?
5630	2496	?
6225	4244	?

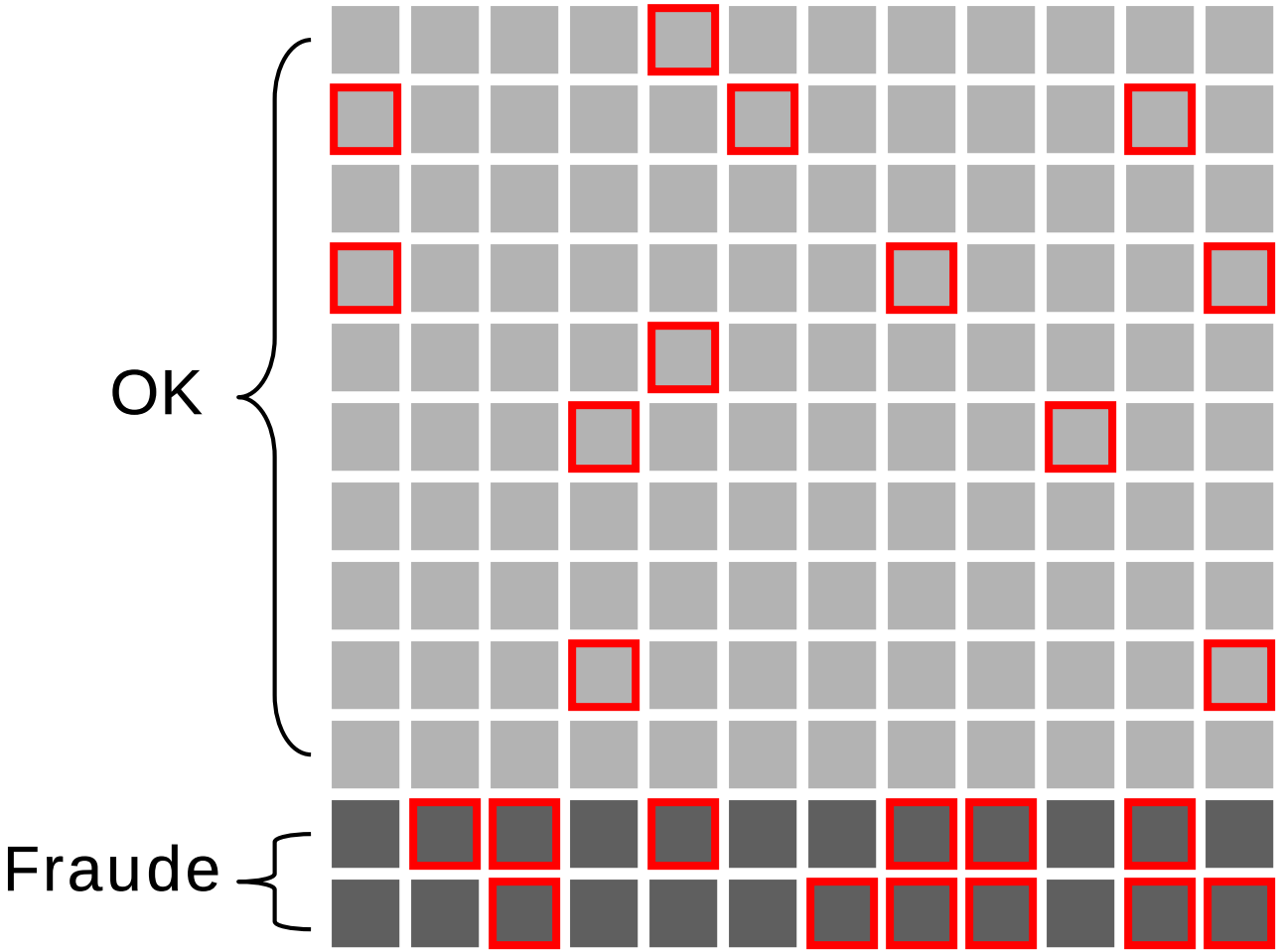
Agrupamiento Vertical

Frequent Flier	Month	Flying Mileage	VIP Member
10621	Jan	650	No
10621	Feb	0	No
10621	Mar	0	No
10621	Apr	250	No
33855	Jan	350	No
33855	Feb	300	No
33855	Mar	1200	Yes
33855	Apr	850	Yes

La difícil búsqueda del objetivo



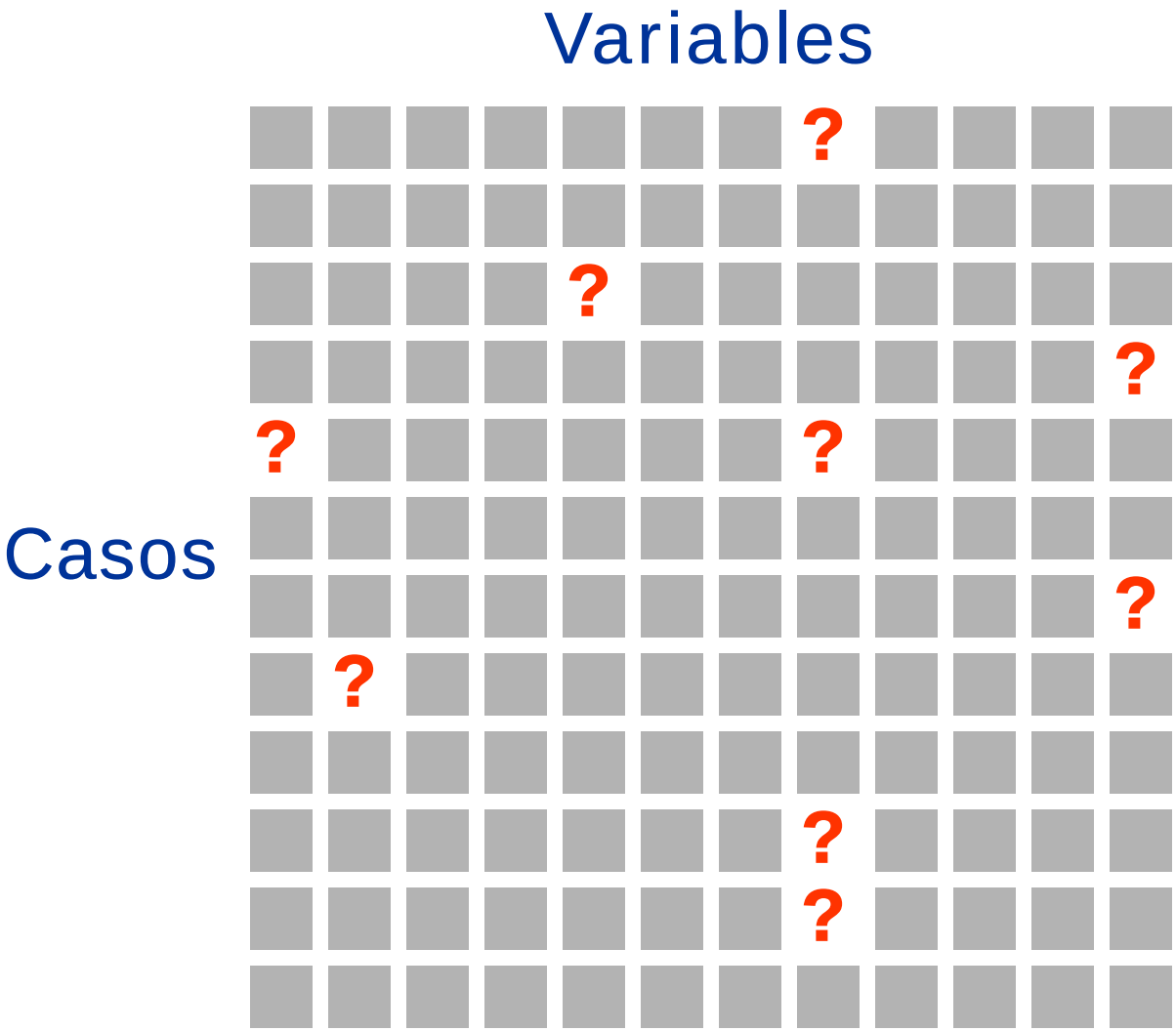
Sobremuestreo (Oversampling)



Errores, Valores Extremos y Valores Perdidos

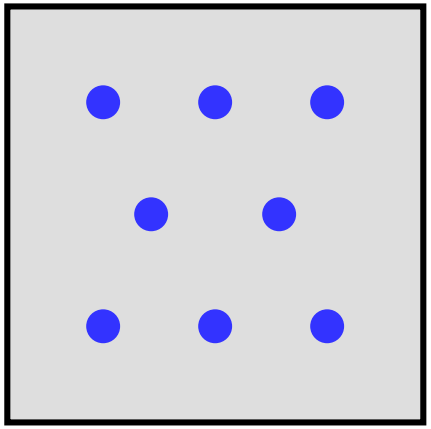
cking	#cking	ADB	NSF	dirdep	SVG	bal
Y	1	468.11	1	1876	Y	1208
Y	1	68.75	0	0	Y	0
Y	1	212.04	0	6		0
	.	.	0	0	Y	4301
y	2	585.05	0	7218	Y	234
Y	1	47.69	2	1256		238
Y	1	4687.7	0	0		0
	.	.	1	0	Y	1208
Y	.	.	.	1598		0
	1	0.00	0	0		0
Y	3	89981.12	0	0	Y	45662
Y	2	585.05	0	7218	Y	234

Asignación de Valores a los Valores Perdidos

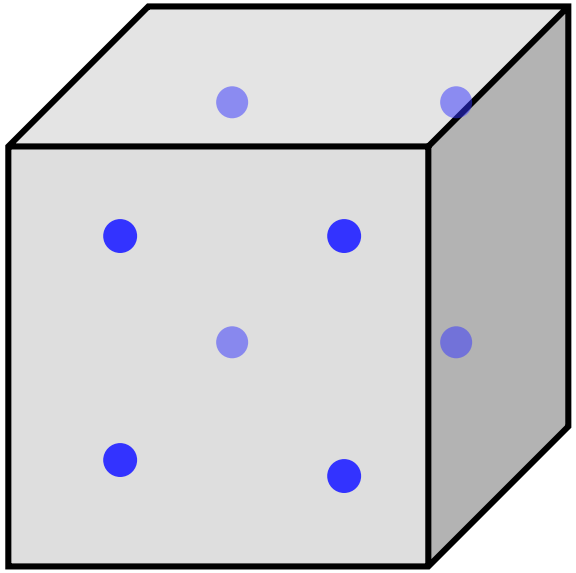


El Problema de la Dimensionalidad

1–D



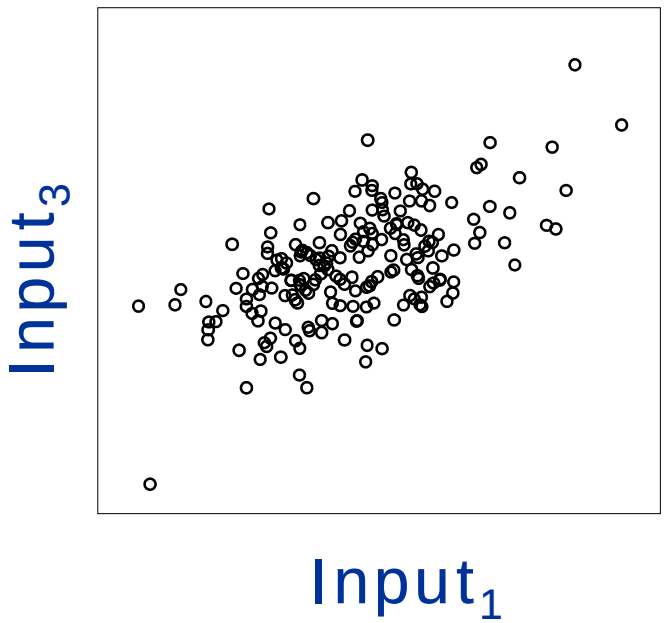
2–D



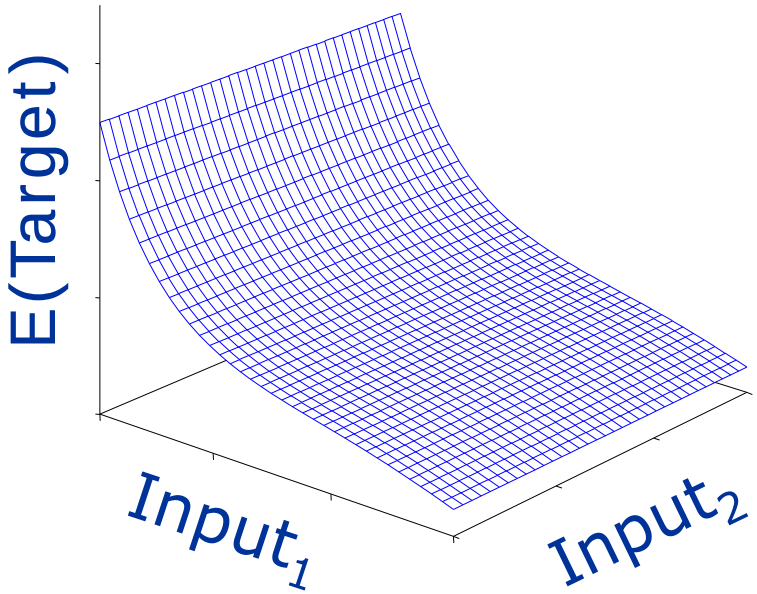
3–D

La Reducción de la Dimensión

Redundancia

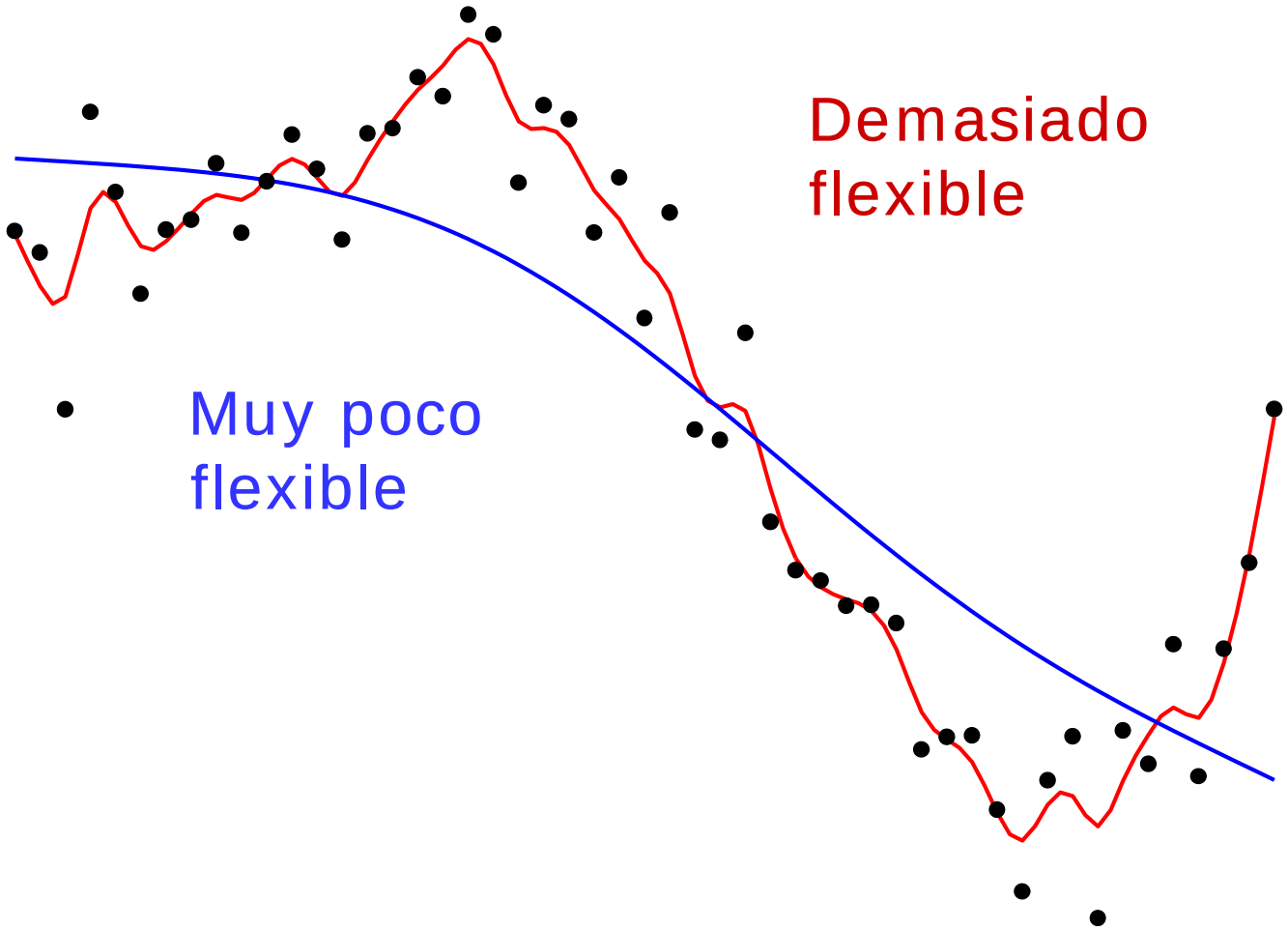


Irrelevancia



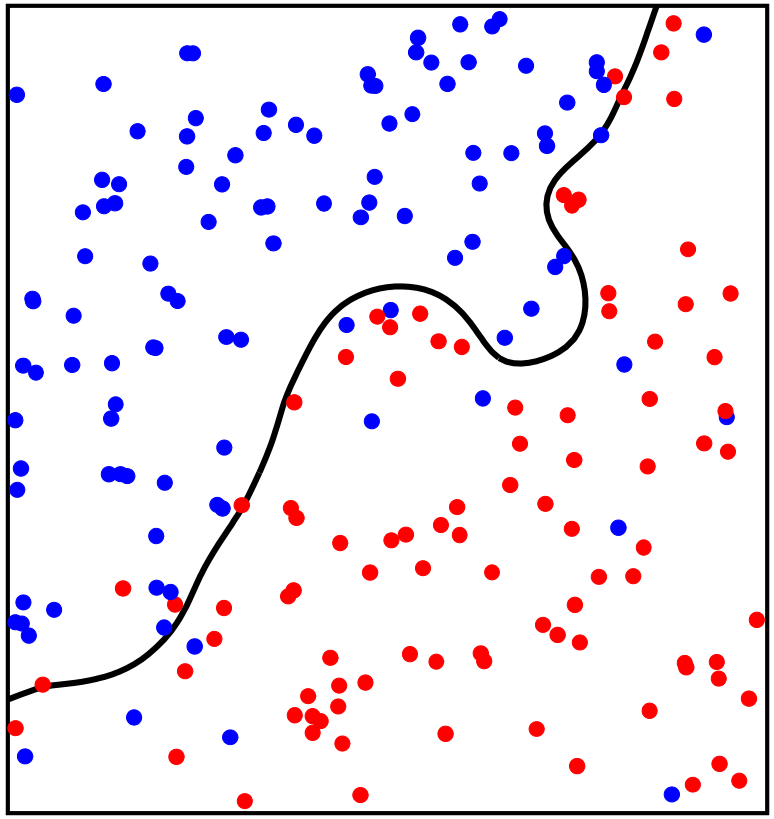
La Partición de los Datos

La Complejidad del Modelo

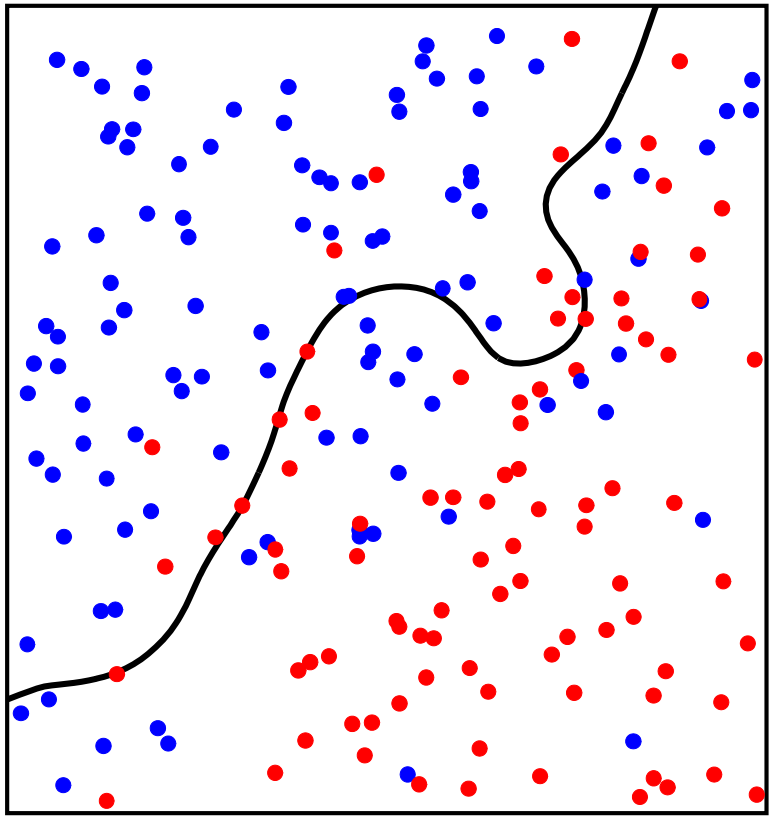


Sobreajuste (Overfitting)

Training Set

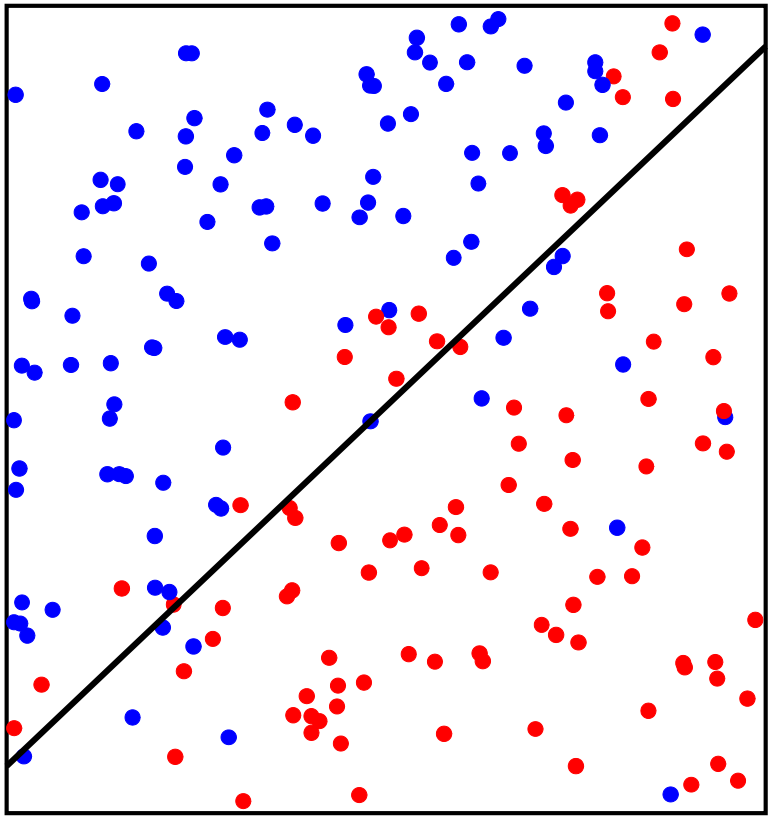


Test Set

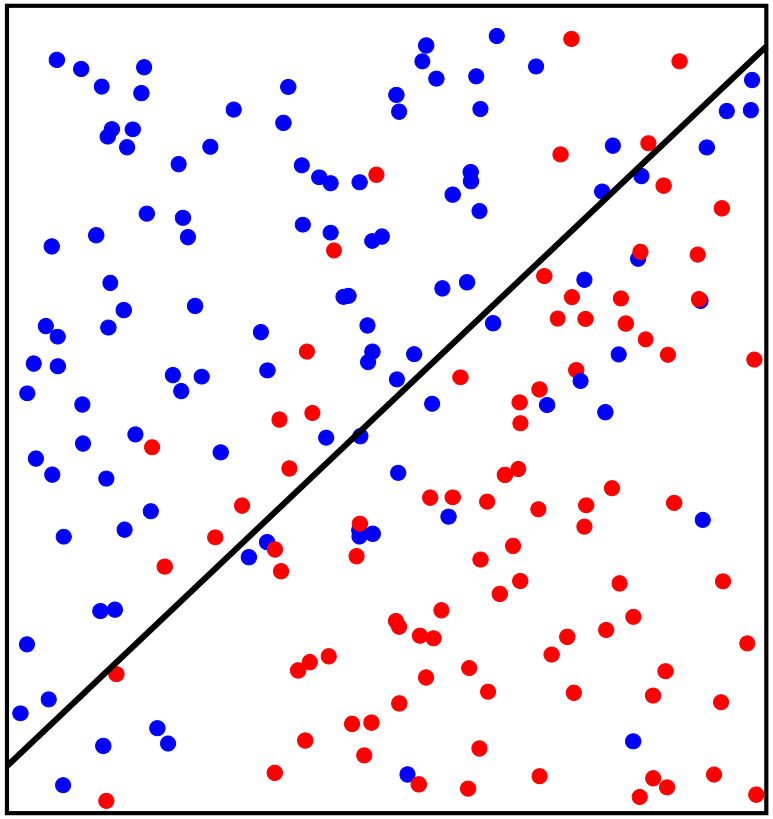


Buen Ajuste

Training Set

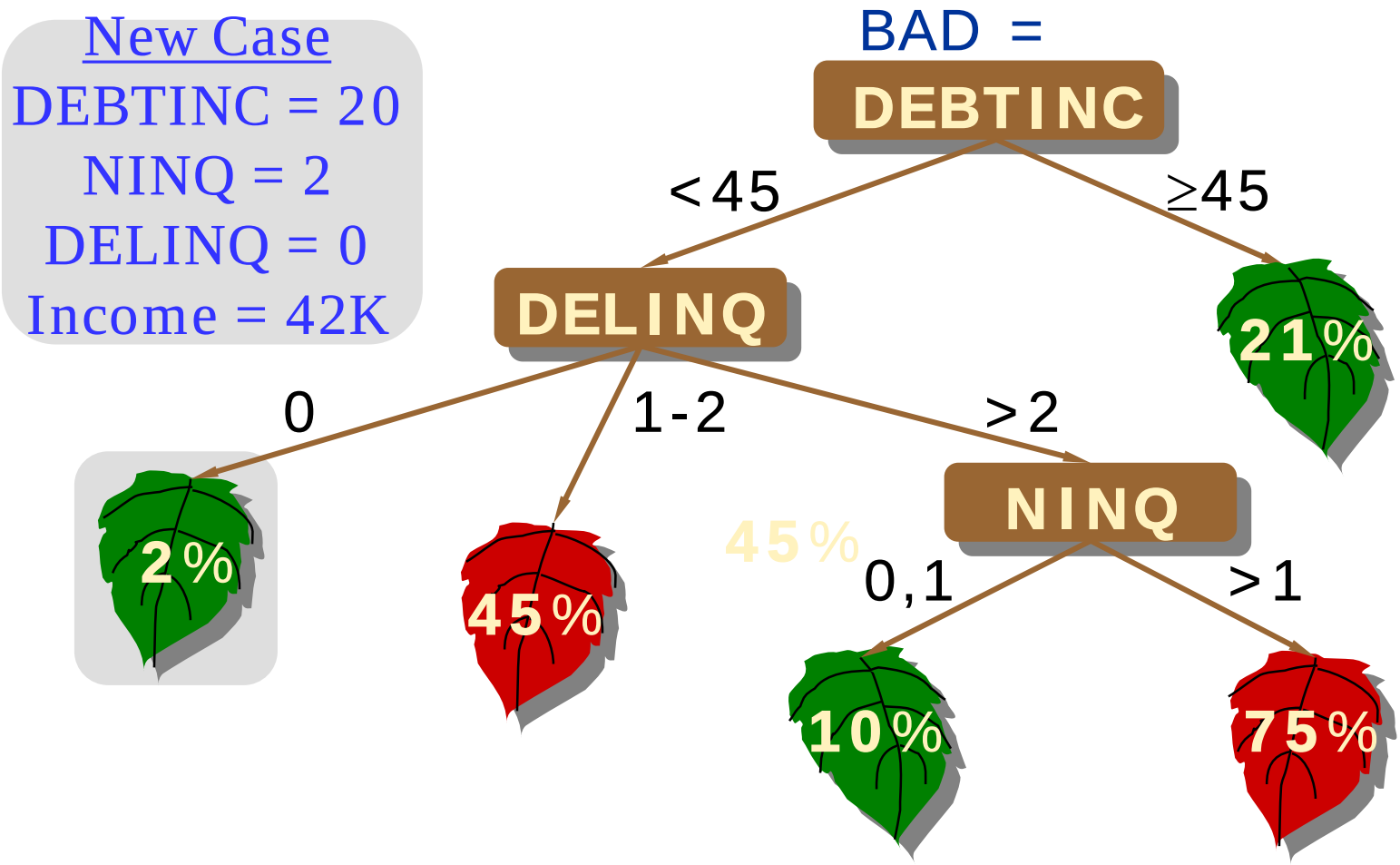


Test Set

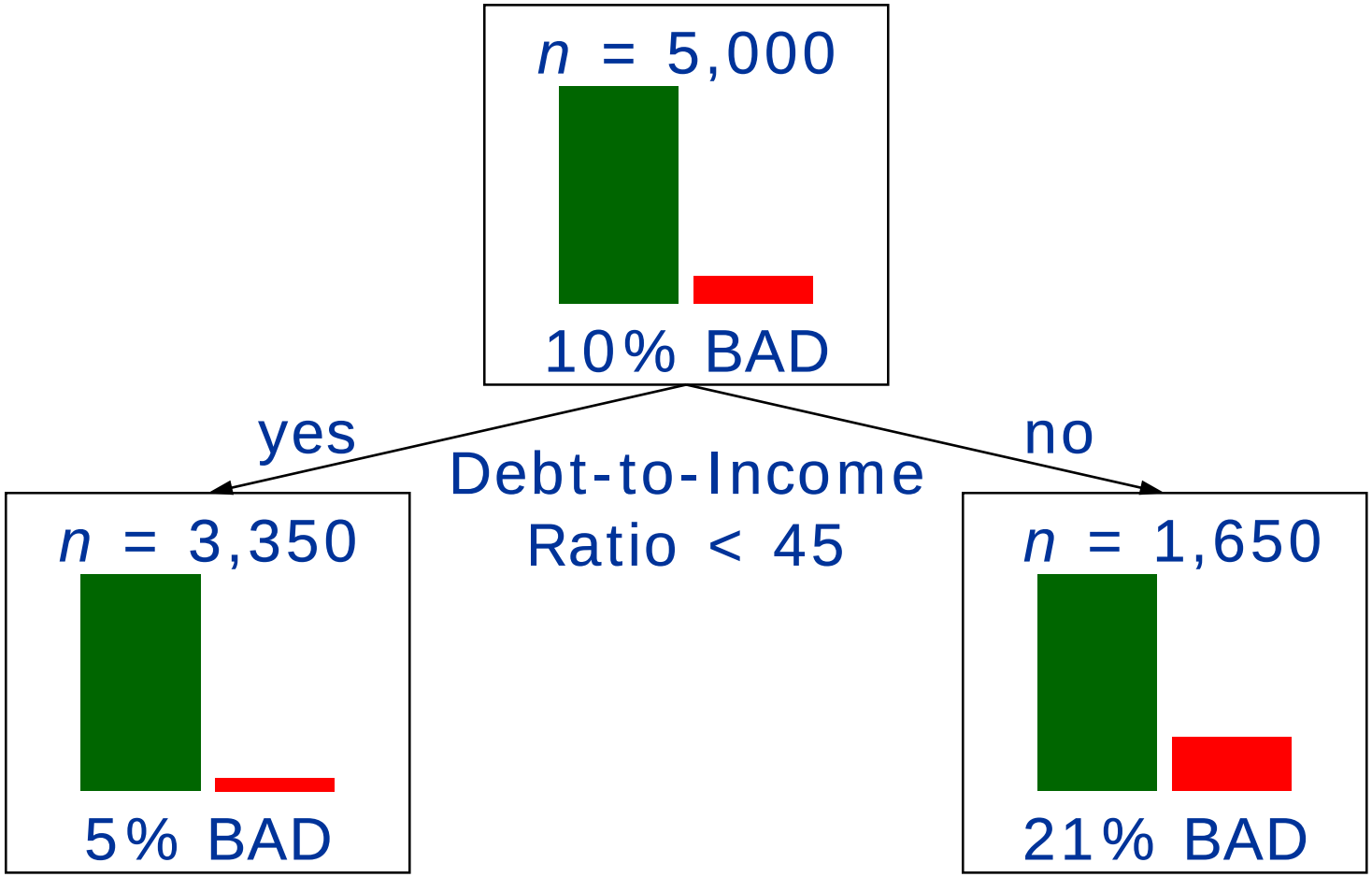


Introducción a la Modelación Predictiva Utilizando Árboles de Decisión

Modelo de Árbol de Decisión



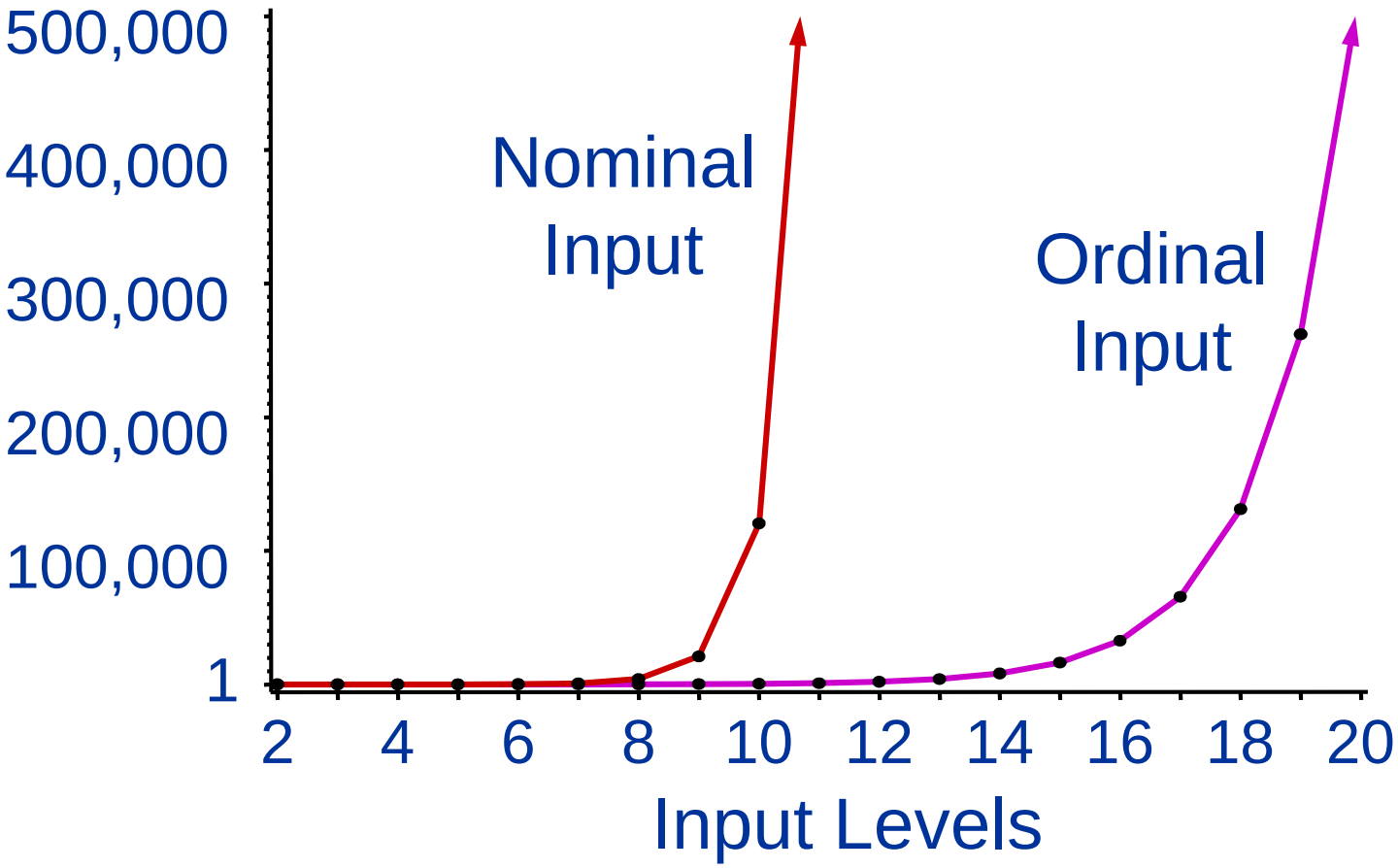
Divide y Vencerás



El Cultivo de los Árboles

- Búsqueda de las particiones
  - ¿Qué particiones deben ser considerada?
- Criterio de particionamiento
  - ¿Qué partición es la mejor?
- Regla de parada
  - ¿Cuándo se deben parar las particiones?
- La regla de poda
  - ¿Se deberán cortar algunas ramificaciones?

Posibles Particiones a Considerar



Criterio de Partición

	Left	Right	
Not Bad	3196	1304	4500
Bad	154	346	500

Debt-to-Income Ratio < 45

	Left	Center	Right	
Not Bad	2521	1188	791	4500
Bad	115	162	223	500

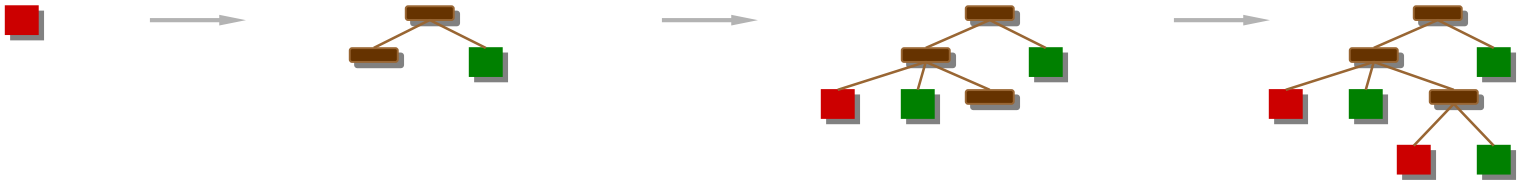
A Competing Three-Way Split

Not Bad	4500	0	4500
Bad	0	500	500

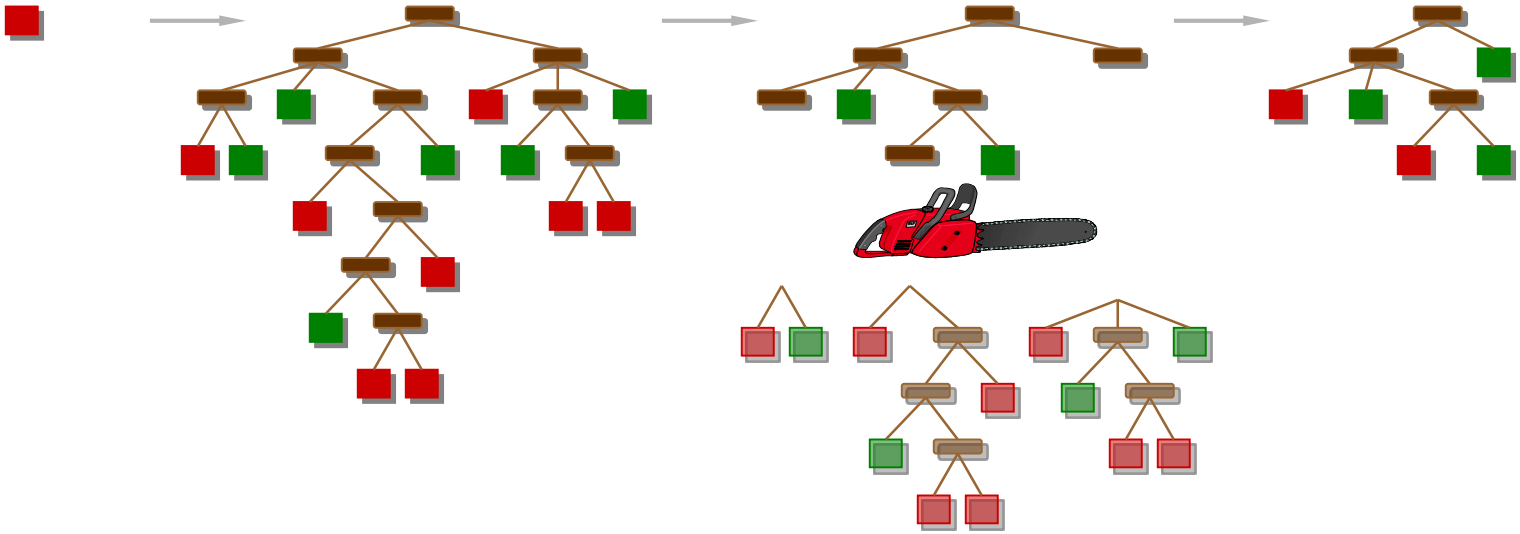
Perfect Split

El Tamaño Correcto del Árbol

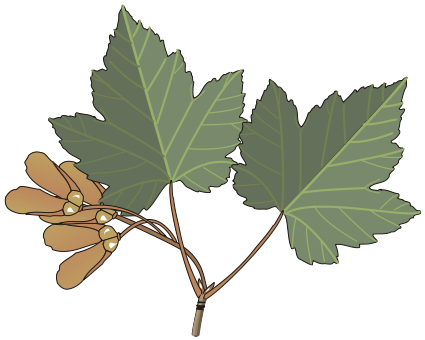
Limitar el crecimiento



Podar

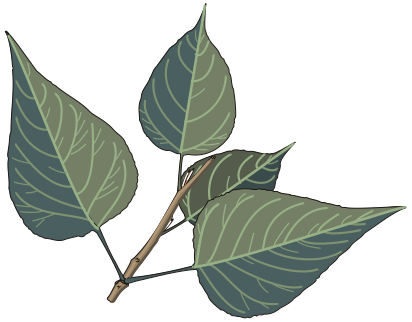


Una Guía de Campo para los Algoritmos de Árboles



AID
THAID
CHAID

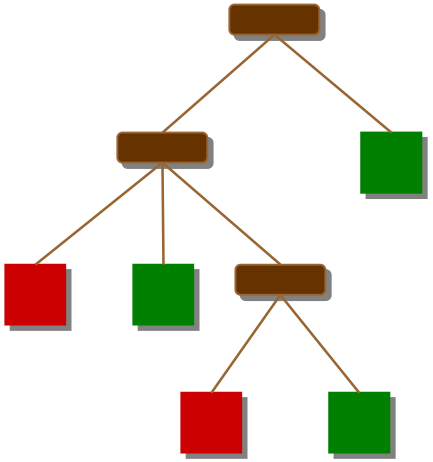
CART



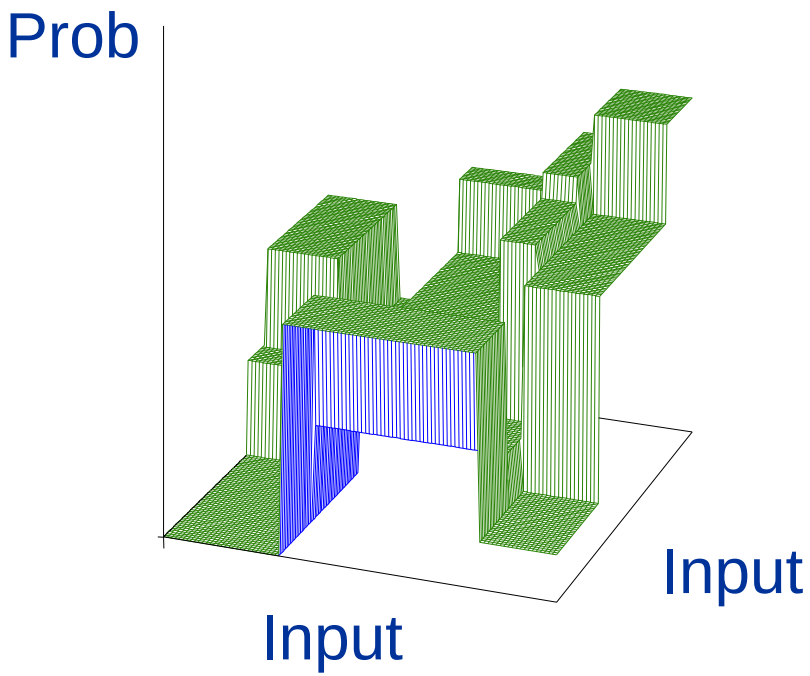
ID3
C4.5
C5.0

Beneficios de los Árboles

- Interpretabilidad.
  - Presentación estructurada.
- Mezcla de escalas de medición.
  - nominal, ordinal, intervalos
- Árboles de regresión.
- Robustez.
- Buen manejo de los valores perdidos.



Beneficios de los Árboles

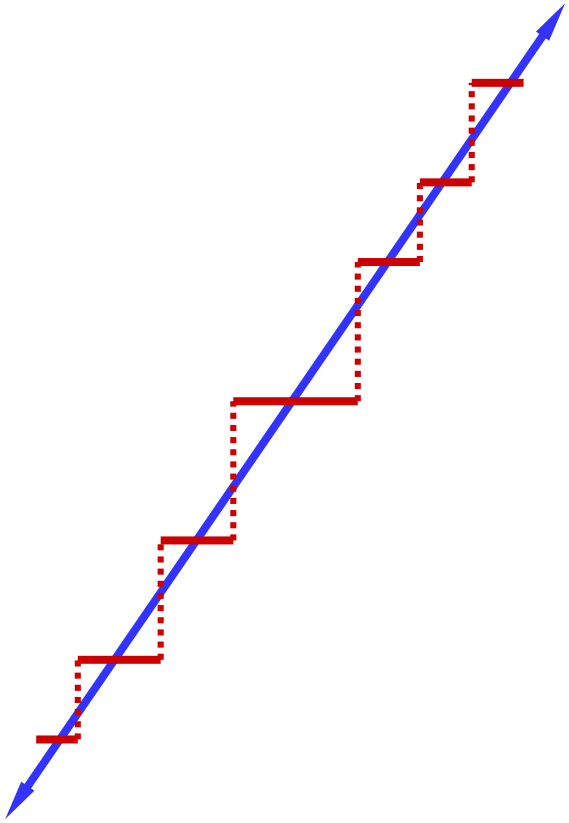


Multivariate
Step Function

- Automáticamente:
  - Detecta las interacciones (AID).
  - Acomoda las no linealidades.
  - Selecciona las variables de entrada en orden jerárquico.

Inconvenientes de los Árboles

- Aspereza (particiones duras).
- Lineal (aditividad de efectos lineales en el ajuste).
- Inestabilidad (pequeñas perturbaciones producen efectos importantes en la topología).



Demostración

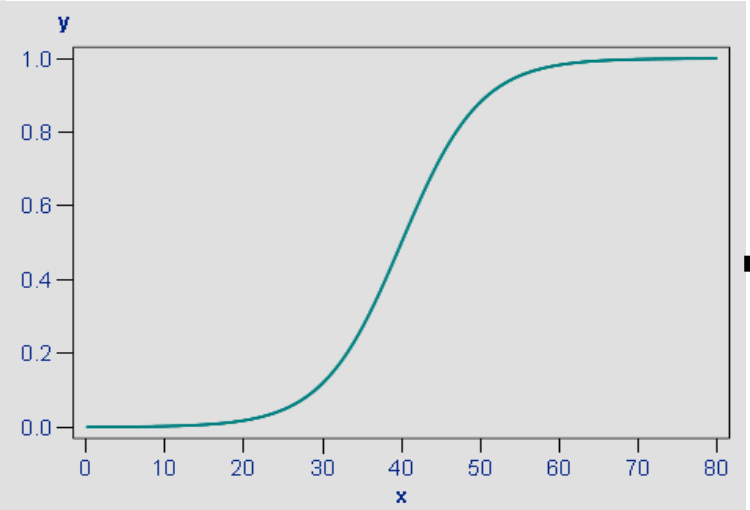
Esta demostración ilustra la construcción de modelo de árbol con el minero de datos y el análisis de los resultados.

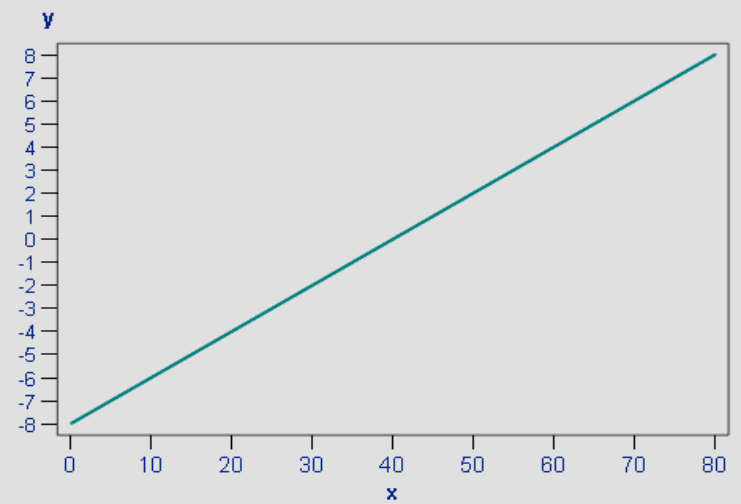
Introducción a la Modelación Predictiva Utilizando Regresión Logística

Regresión Lineal versus La Logística

Regresión Lineal	Regresión Logística
La target es una variable continua (intervalo).	La target es una variable discreta (binaria u ordinal).
Las variables de entrada tienen varios niveles de medida.	Las variables de entrada tienen varios niveles de medida.
Los valores predichos son la media de la target para valores dados de las variables de entrada.	Los valores predichos son la probabilidad de niveles particulares de la target para valores dados de las variables de entrada.

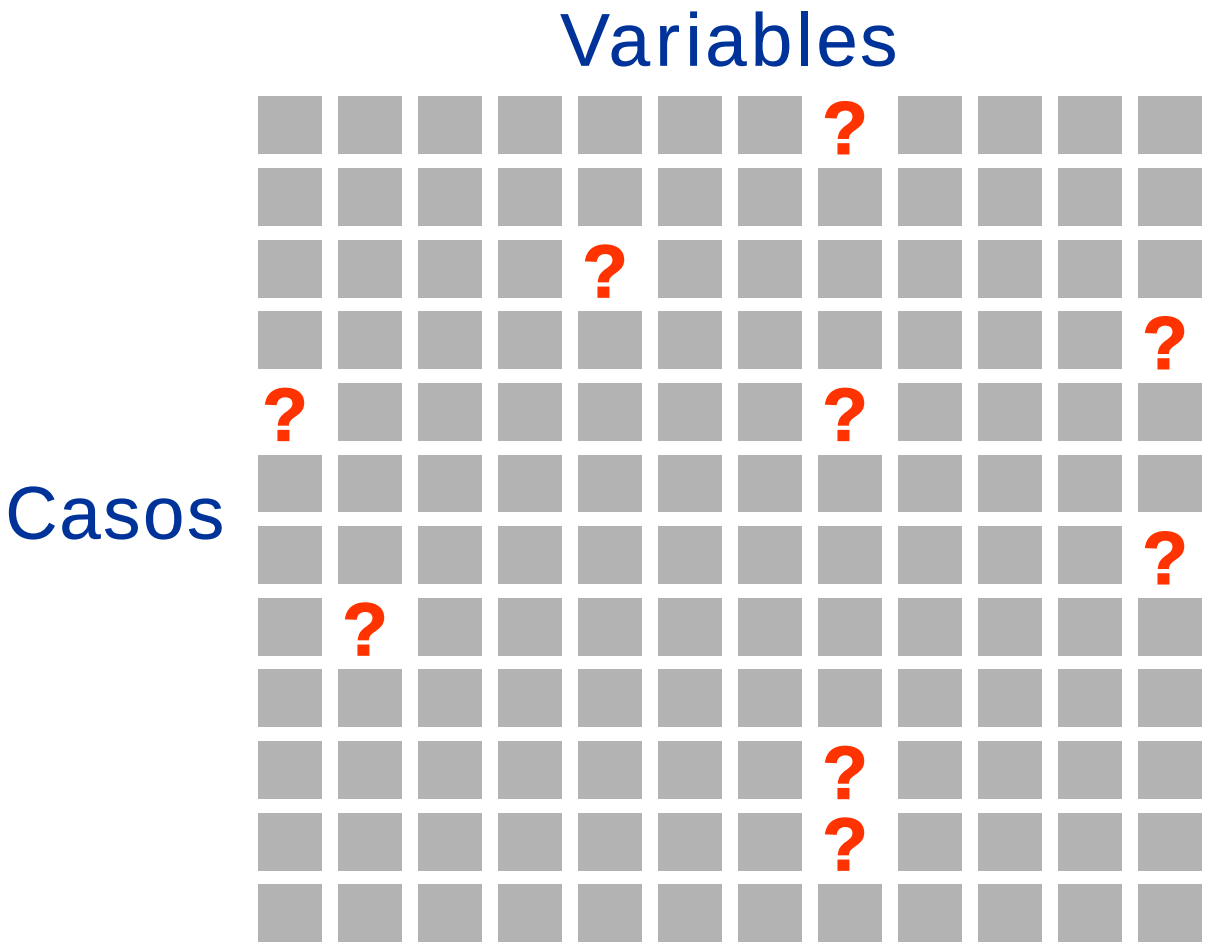
Suposición de la Regresión Logística



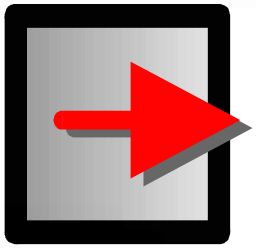


logit transformation

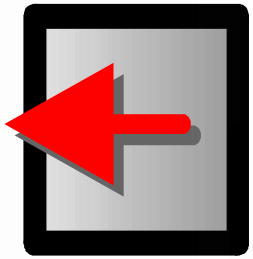
Los Valores Perdidos



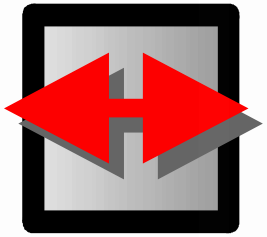
Métodos de Regresión



Forward Selection



Backward Selection



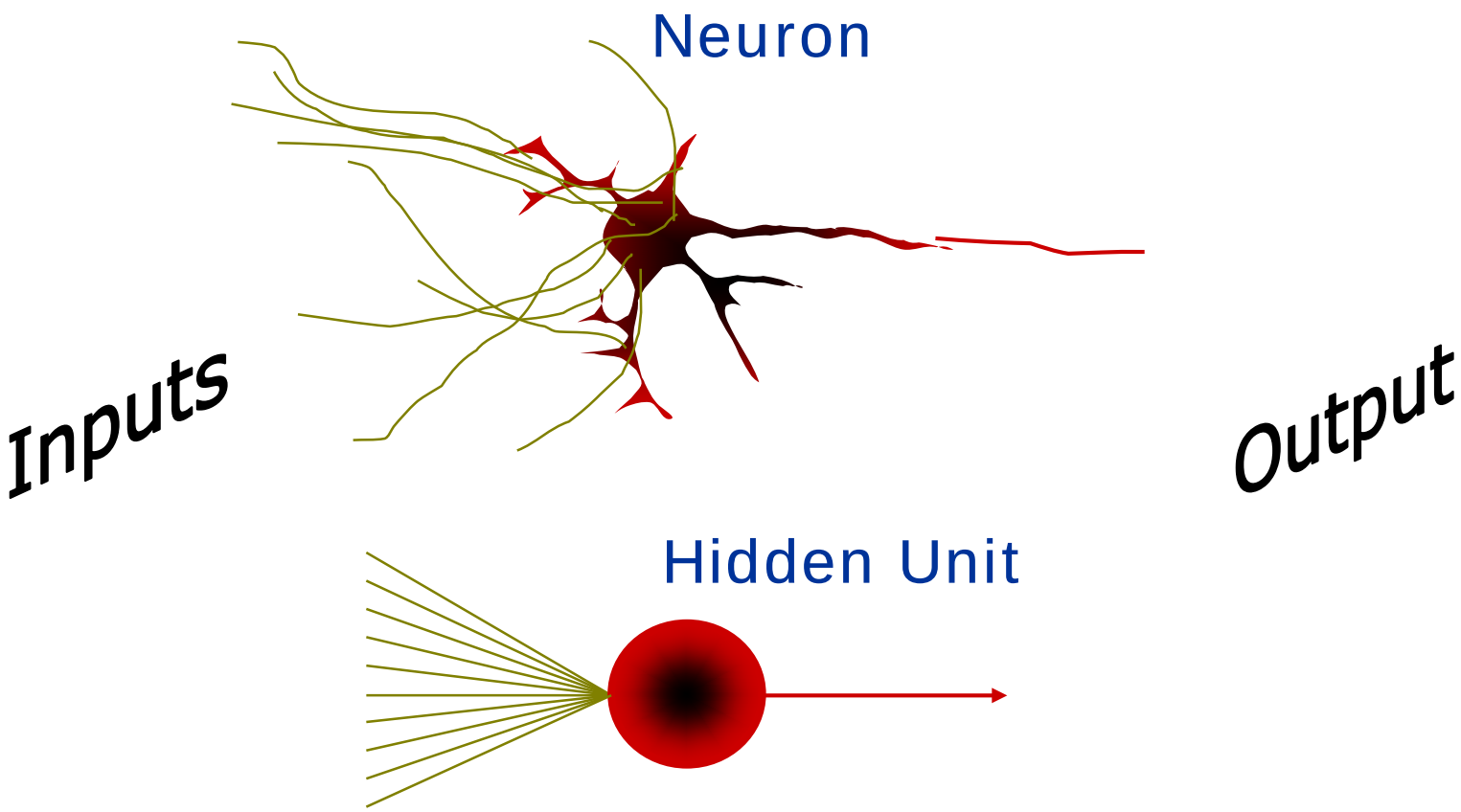
Stepwise Selection

Demostración

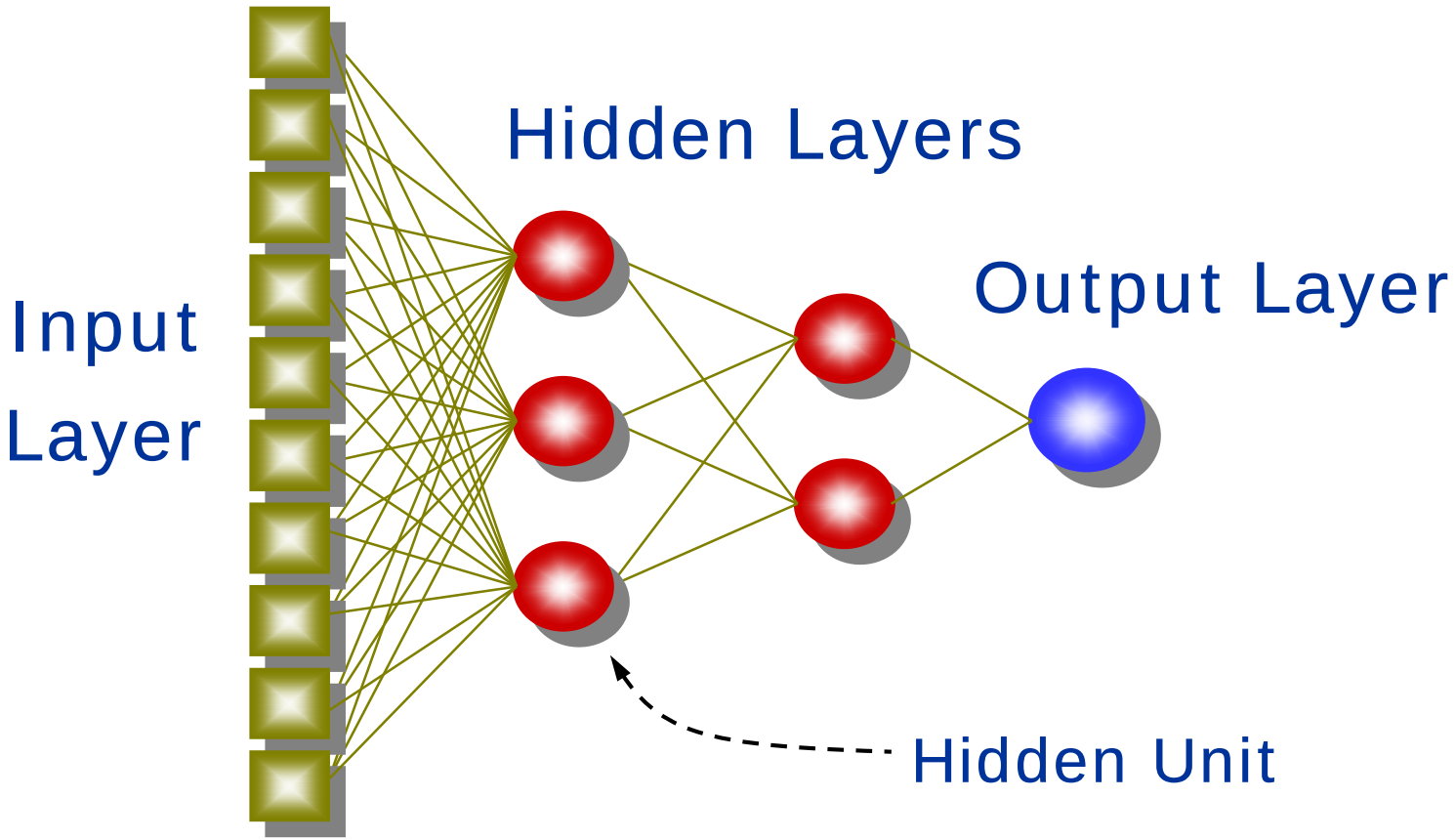
Esta demostración ilustra la asignación de valores a los valores perdidos y la generación de un modelo de regresión.

Introducción a la Modelación Predictiva Utilizando Redes Neuronales

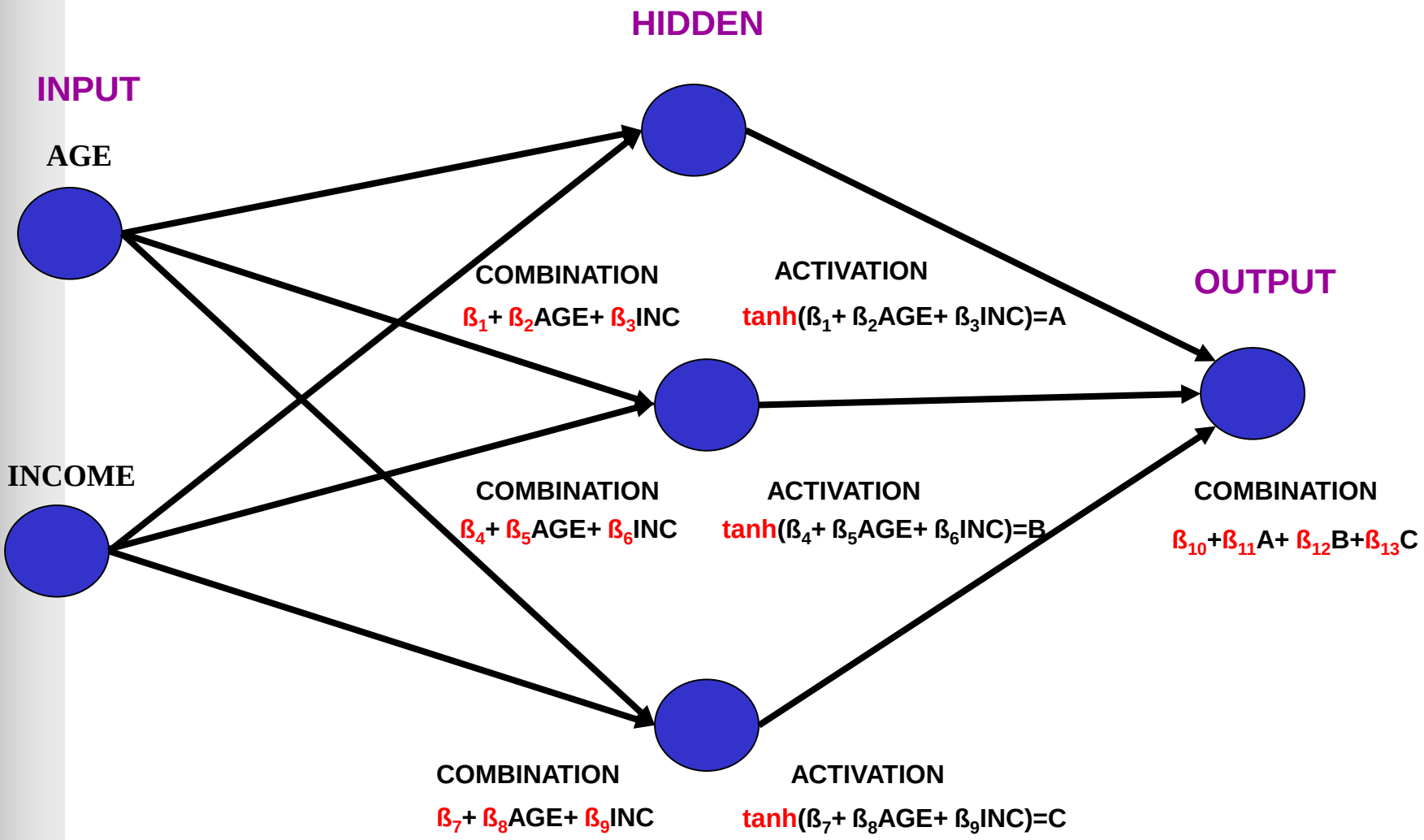
Red Neuronal Artificial



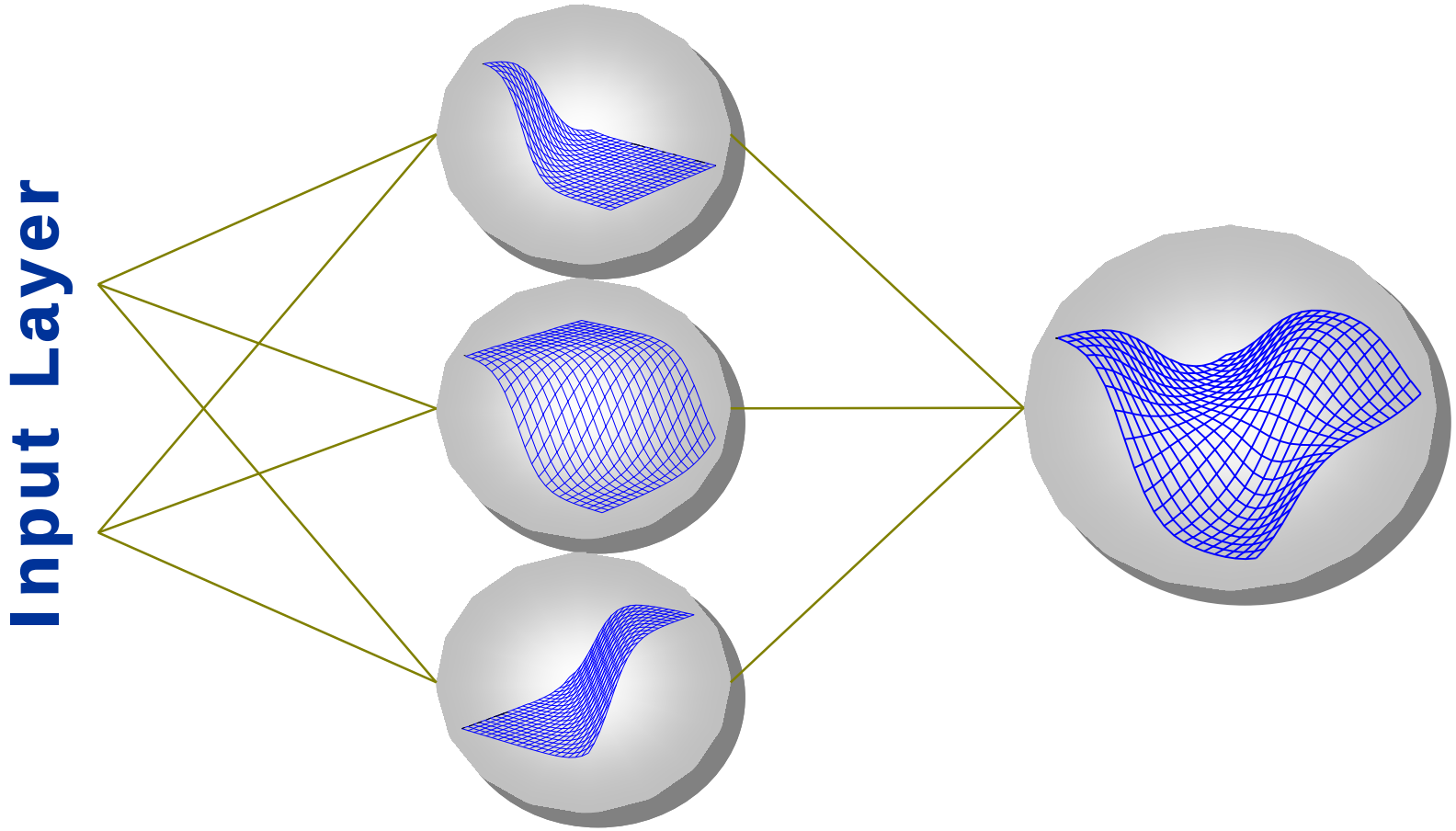
Multilayer Perceptron



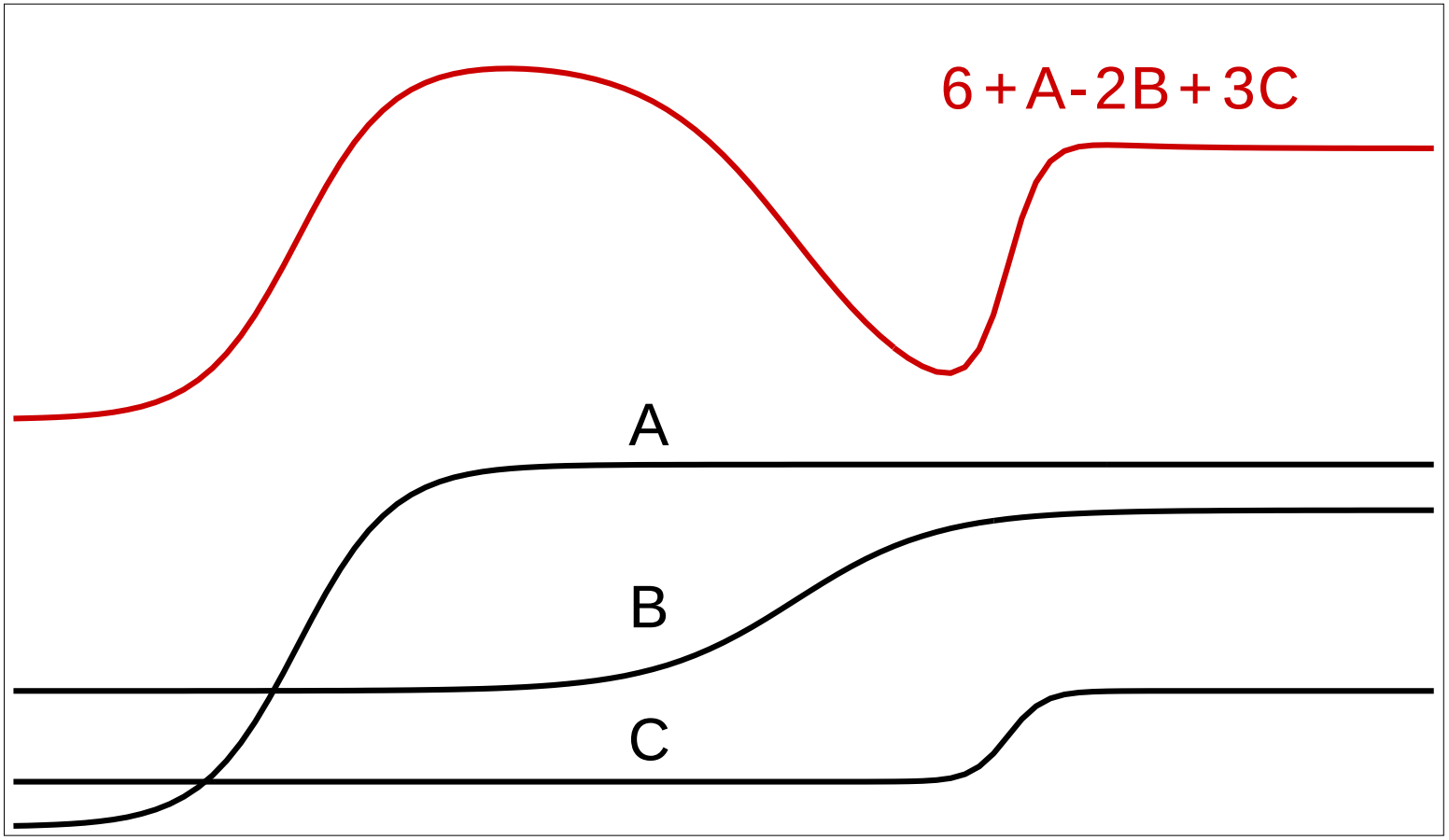
Estructura del Modelo



Función de Activación

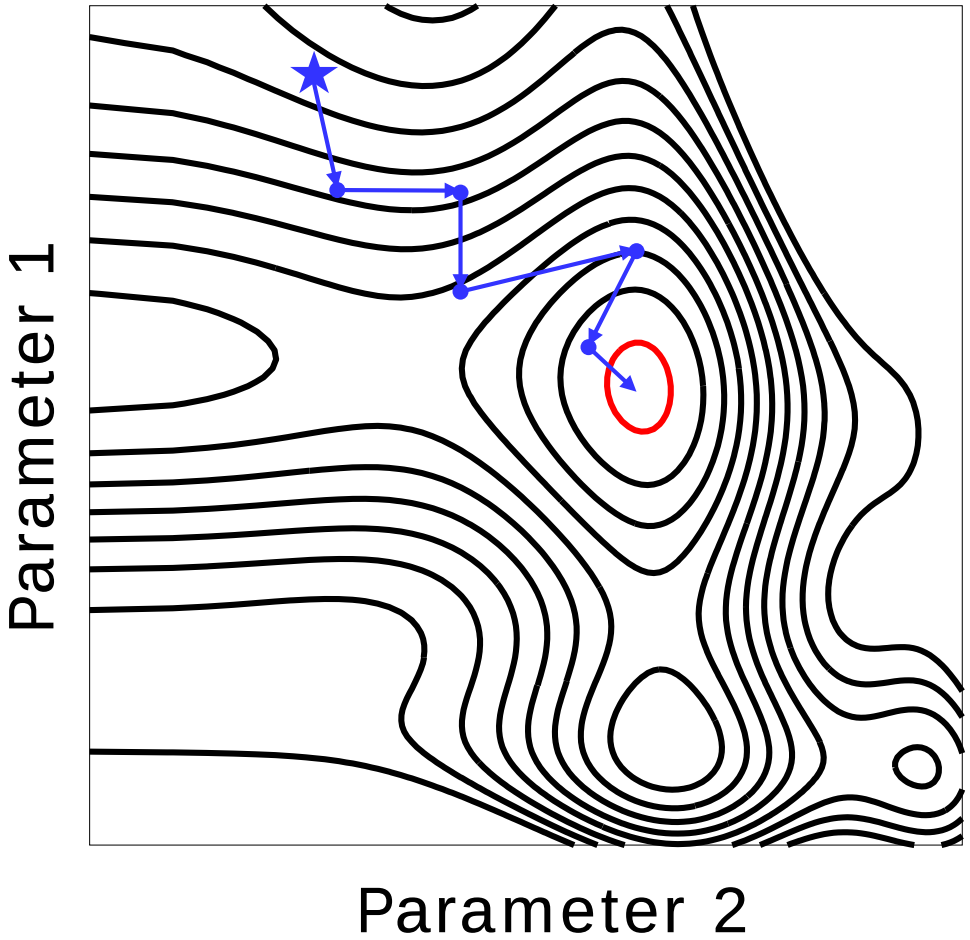


Aproximador Universal



Entrenamiento de la Red

- Función de error.
- Algoritmo de optimización iterativo. Método del Ascenso Acelerado.



Demostración

Esta demostración ilustra el ajuste de un modelo de red neuronal.