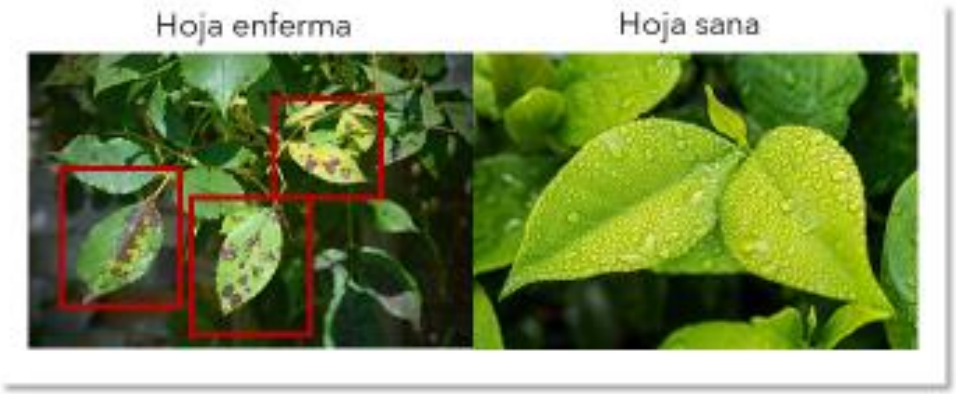
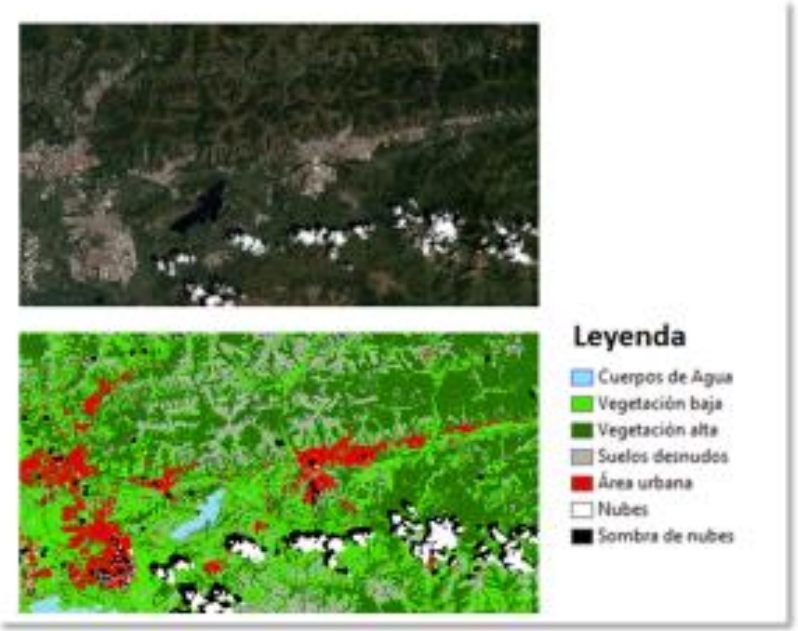
Sesión 2. Clasificación. Regresión logística

- La tarea de clasificación.
- Discriminantes lineales.
- El algoritmo de regresión logística.

¿Cómo resolver una tarea de clasificación?





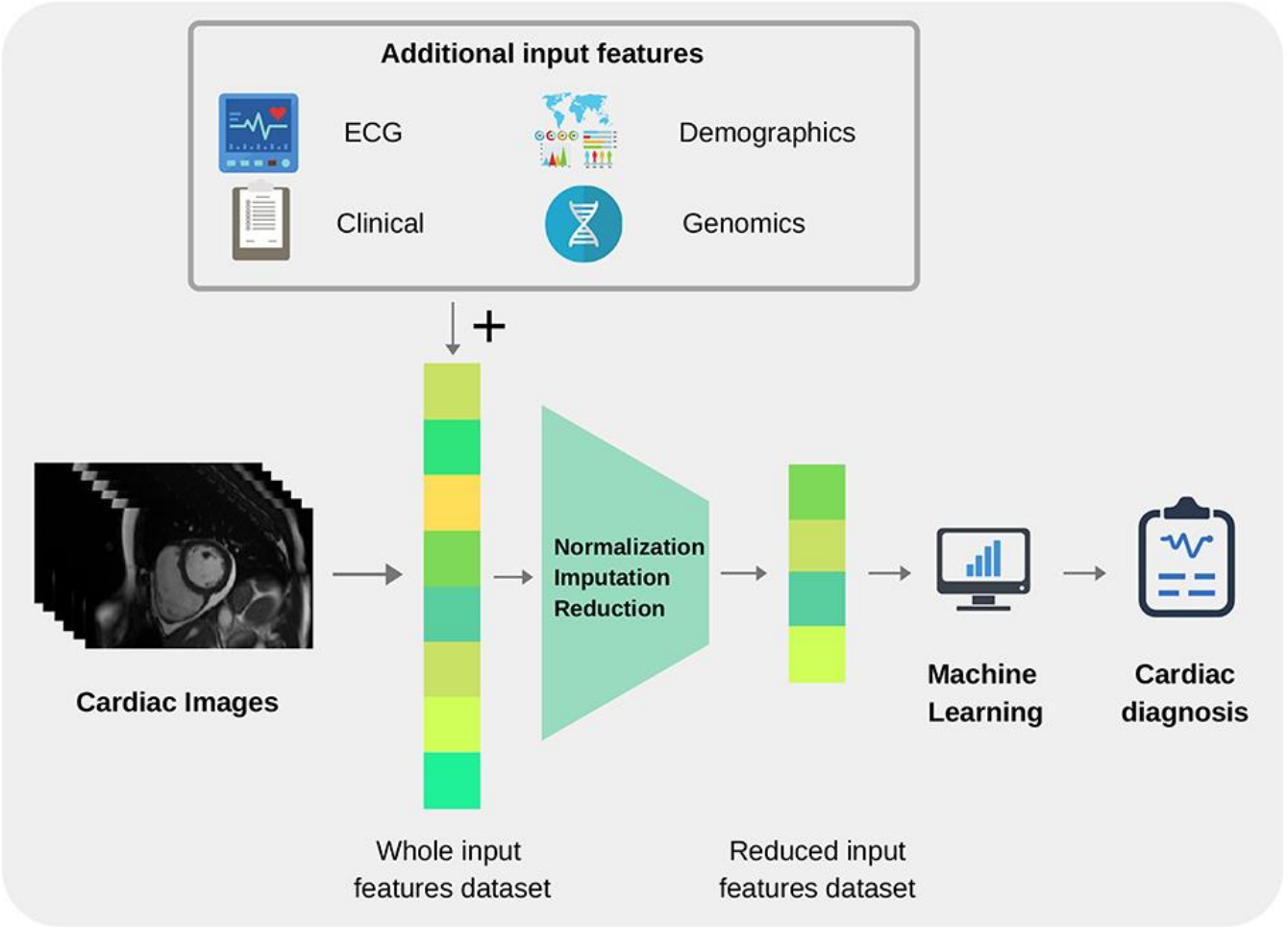
ML en salud

¿Datos?

- Apoyo al diagnóstico de enfermedades.
- Interpretación de imágenes.
- Medicina personalizada.
- Planificación de radioterapia.
- Detección de anomalías en señales de equipos médicos.
- Gestión hospitalaria, ….

Ejemplo:

Image-Based Cardiac Diagnosis With Machine Learning: A Review



https://www.frontiersin.org/articles/10.3389/fcvm.2020.00001/full

Proyecto: estudio sobre las enfermedades cardiovasculares con técnicas de machine learning.

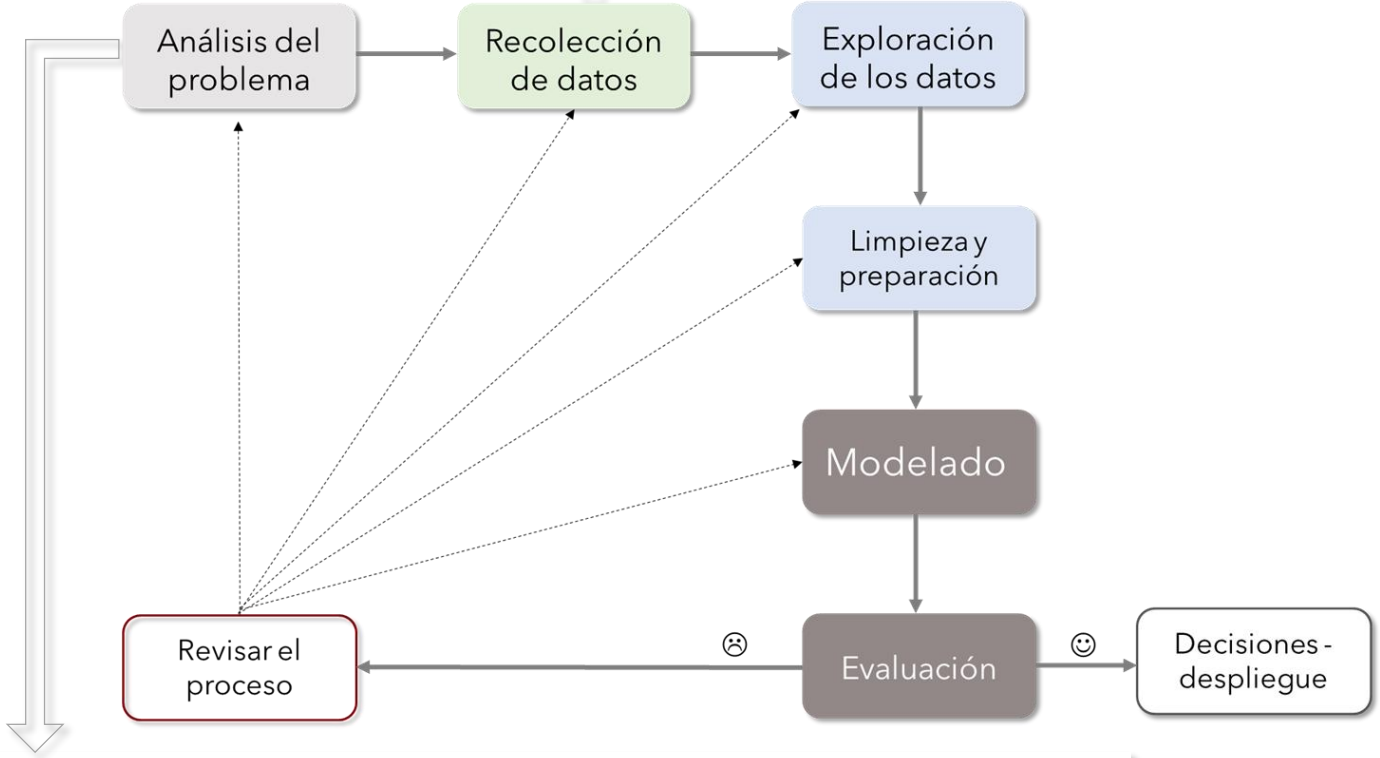
Según la Organización Mundial de la Salud , las enfermedades cardiovasculares (ECV) son la principal causa de muerte en todo el mundo, y se estima que, para el 2030, 23,6 millones de personas morirán por alguna enfermedad de este tipo. Las ECV son un grupo de trastornos del corazón y de los vasos sanguíneos que incluyen, entre otras afecciones, a la enfermedad coronaria, enfermedad cerebrovascular y enfermedad cardíaca reumática. Cuatro de cada 5 muertes por ECV se deben a ataques cardíacos y accidentes cerebrovasculares, y un tercio de estas muertes ocurren prematuramente en personas menores de 70 años. Un pronóstico temprano podría ayudar a tomar decisiones sobre los cambios en el estilo de vida en pacientes de alto riesgo y, a su vez, reducir las complicaciones.
En este sentido, se quiere llevar a cabo un estudio, tomando como base los datos de los pacientes registrados en una institución médica, que utilice técnicas de machine learning para la construcción de modelos que, no solo aporten en la identificación de los factores que más inciden en el padecimiento de enfermedad cardíaca, sino también puedan predecir qué pacientes están en riesgo de sufrir una ECV.

age	gender	height	weight	systolic blood pressure	diastolic blood pressure	cholesterol	glucose	smoke	Alcohol	active	ECV
18393	2	168	62.0	110	80	normal	normal	0	0	1	0
20228	1	156	85.0	140	90	well above normal	normal	0	0	1	1
18857	1	165	64.0	130	70	well above normal	normal	0	0	0	1
17623	2	169	82.0	150	100	normal	normal	0	0	1	1
17474	1	156	56.0	100	60	normal	normal	0	0	0	0
21914	1	151	67.0	120	80	above normal	above normal	0	0	0	0
22113	1	157	93.0	130	80	well above normal	normal	0	0	1	0
22584	2	178	95.0	130	90	well above normal	well above normal	0	0	1	1
17668	1	158	71.0	110	70	normal	normal	0	0	1	0
19834	1	164	68.0	110	60	normal	normal	0	0	0	0
22530	1	169	80.0	120	80	normal	normal	0	0	1	0

Número de datos: 69.997
Número de variables: 12

https://www.who.int/es/health-topics/cardiovascular-diseases#tab=tab_1

Ciclo de ML



Primera visión de los datos:

Ayuda a determinar si el conjunto de datos es adecuado para resolver el problema.

Del análisis de los datos

Del análisis del problema

Construcción de un modelo de clasificación

Solución propuesta: modelo de clasificación que permita predecir si un paciente tiene posibilidad de ser diagnosticado con ECV.

Objetivos:

- Predecir qué pacientes están en riesgo de sufrir una ECV
- Identificar los factores que más inciden en el padecimiento de enfermedad cardíaca.

Tareas de aprendizaje:

Para alcanzar el primer objetivo se puede realizar una tarea de clasificación que utilice como variable target “ECV”, la cual indica la presencia o ausencia de enfermedad cardiovascular. Dependiendo del algoritmo a utilizar se podría hacer un análisis de las variables más significativas, lo cual podría permitir alcanzar el segundo objetivo. O utilizar alguna librería que permita realizar una interpretación del modelo.

Variables de entrada = **x**

Clase = y

age	gender	height	weight	systolic blood pressure	diastolic blood pressure	cholesterol	glucose	smoke	Alcohol	active	ECV
18393	2	168	62.0	110	80	normal	normal	0	0	1	0
20228	1	156	85.0	140	90	well above normal	normal	0	0	1	1
18857	1	165	64.0	130	70	well above normal	normal	0	0	0	1
17623	2	169	82.0	150	100	normal	normal	0	0	1	1
17474	1	156	56.0	100	60	normal	normal	0	0	0	0
21914	1	151	67.0	120	80	above normal	above normal	0	0	0	0
22113	1	157	93.0	130	80	well above normal	normal	0	0	1	0
22584	2	178	95.0	130	90	well above normal	well above normal	0	0	1	1
17668	1	158	71.0	110	70	normal	normal	0	0	1	0
19834	1	164	68.0	110	60	normal	normal	0	0	0	0
22530	1	169	80.0	120	80	normal	normal	0	0	1	0
18815	2	173	60.0	120	80	normal	normal	0	0	1	0
14791	2	165	60.0	120	80	normal	normal	0	0	0	0
19809	1	158	78.0	110	70	normal	normal	0	0	1	0

Algoritmo de clasificación

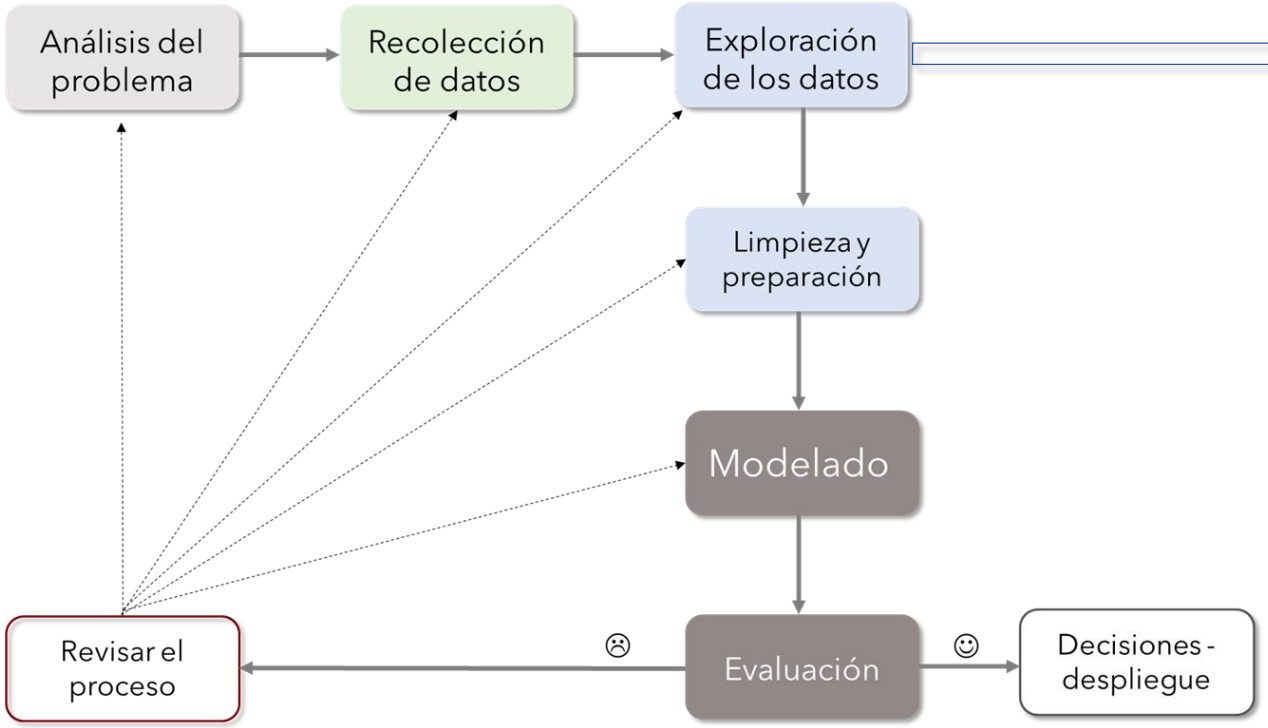
Modelo de clasificación

ECV = SI

ECV = NO

Nuevo dato (paciente)

Edad	Género	Altura	Peso	Presion sistólica	Presión diastólica	Colesterol	Glucosa	Fumador	Alcohol	Activo
50	F	171	76.0	90	60	Normal	Elevado	NO	NO	SI



¿Limpieza de datos?

¿Cómo construir una buena representación de los datos para el algoritmo seleccionado?

¿Qué transformaciones aplicar?

Descripción de los datos (perfilamiento)

Importante: Diccionario de los datos

There are 3 types of input features:

Objective: factual information.

Examination: results of medical examination.

Subjective: information given by the patient.

Features:

Age | Objective Feature | age | int (days)

Height | Objective Feature | height | int (cm) |

Weight | Objective Feature | weight | float (kg) |

Gender | Objective Feature | gender | categorical code |

Systolic blood pressure | Examination Feature | ap_hi | int positive |

Diastolic blood pressure | Examination Feature | ap_lo | int positive |

Cholesterol | Examination Feature | cholesterol | normal, above normal, well above normal |

Glucose | Examination Feature | gluc | normal, above normal, well above normal |

Smoking | Subjective Feature | smoke | binary |

Alcohol intake | Subjective Feature | alco | binary |

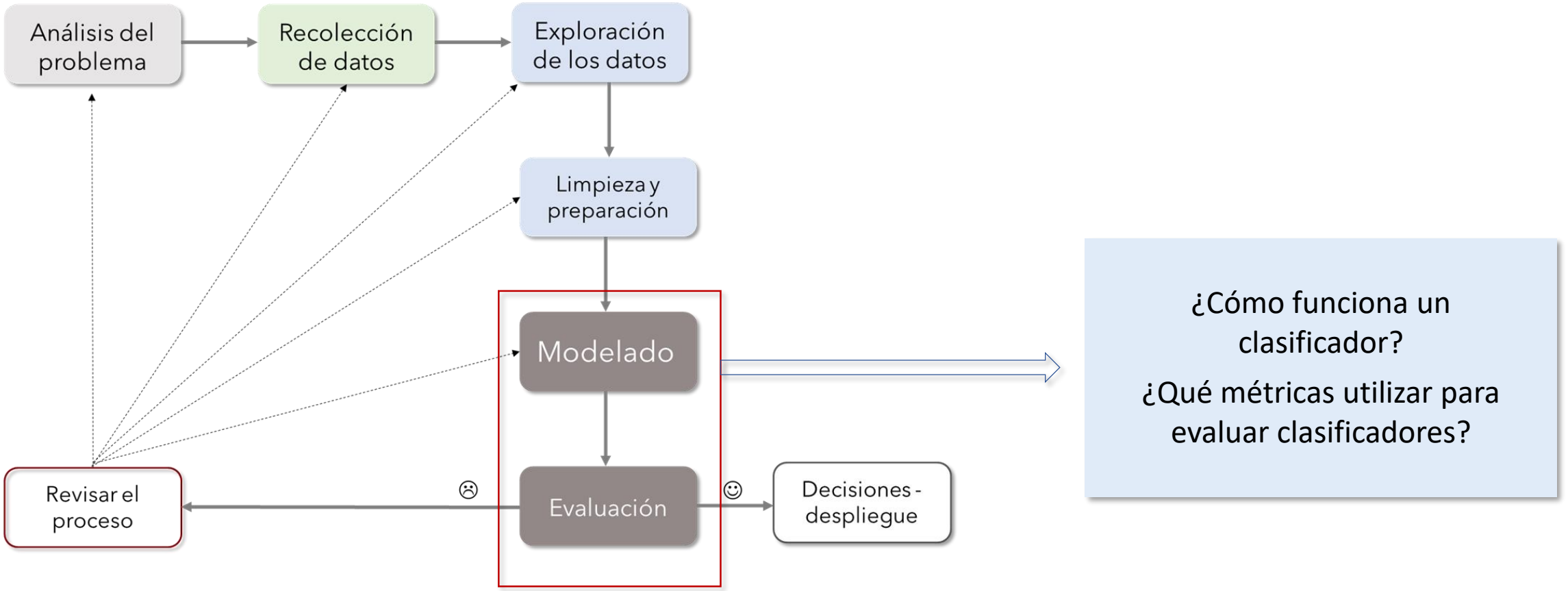
Physical activity | Subjective Feature | active | binary |

Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

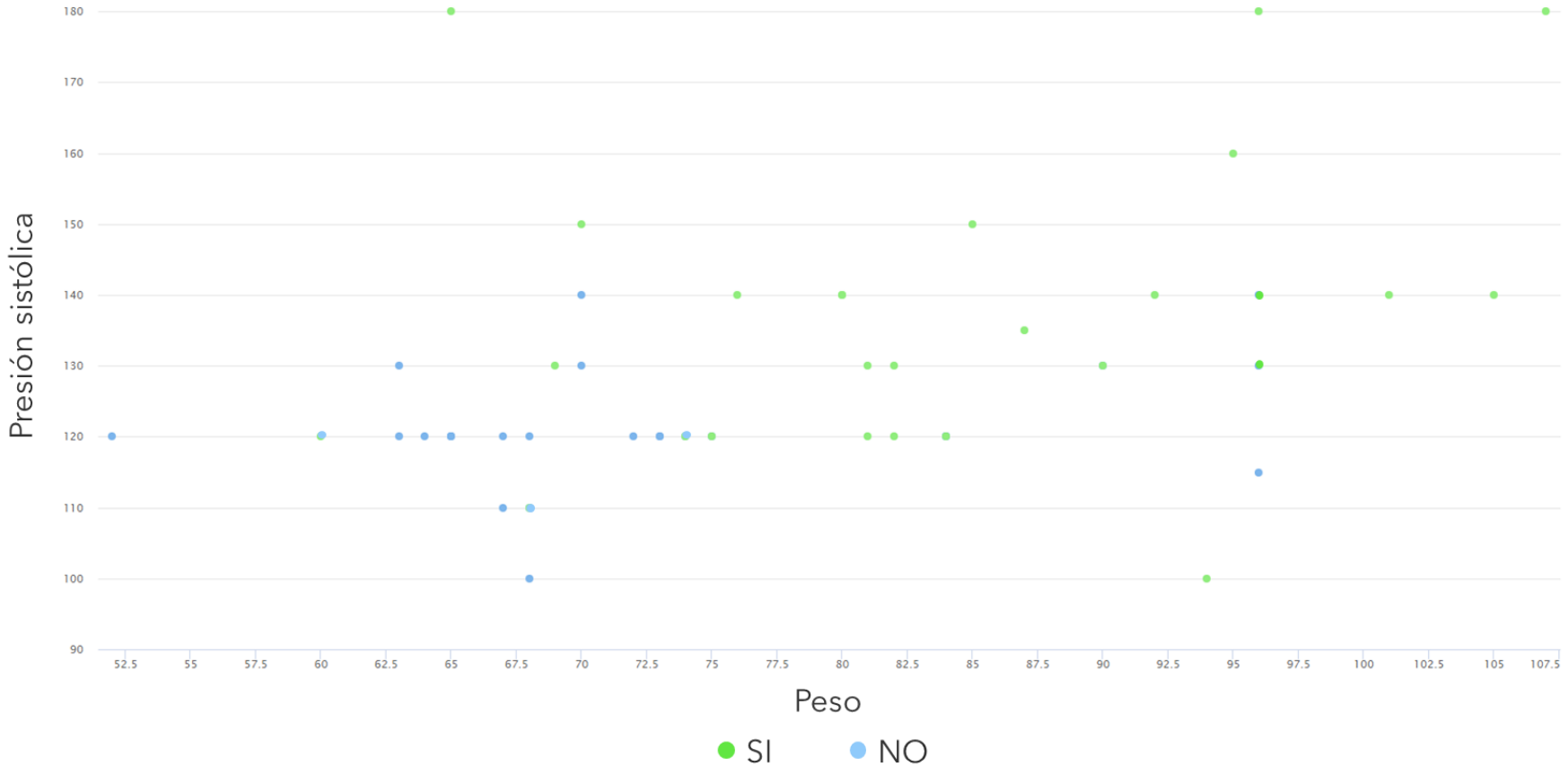
All the dataset values were collected at the moment of medical examination.

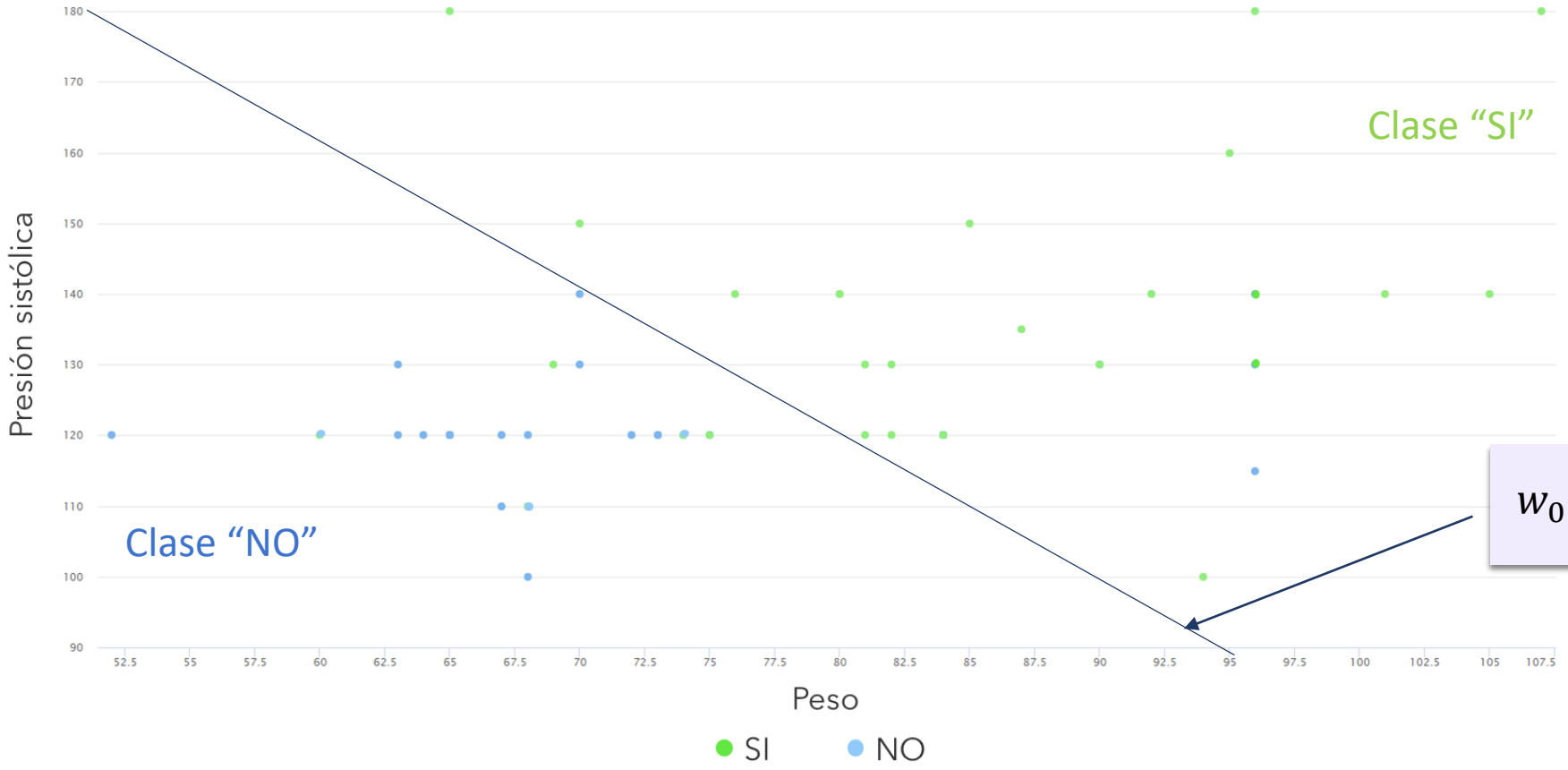
Propuesta para la preparación de los datos:

- Las variables “cholesterol” y “gluc” son categóricas. Como solo asumen tres categorías, se puede aplicar una numerización 1-de-n (one-hot).
- Con base en el algoritmo a utilizar, será conveniente normalizar o estandarizar los datos.



¿Qué hace un algoritmo de aprendizaje para resolver este problema de clasificación?





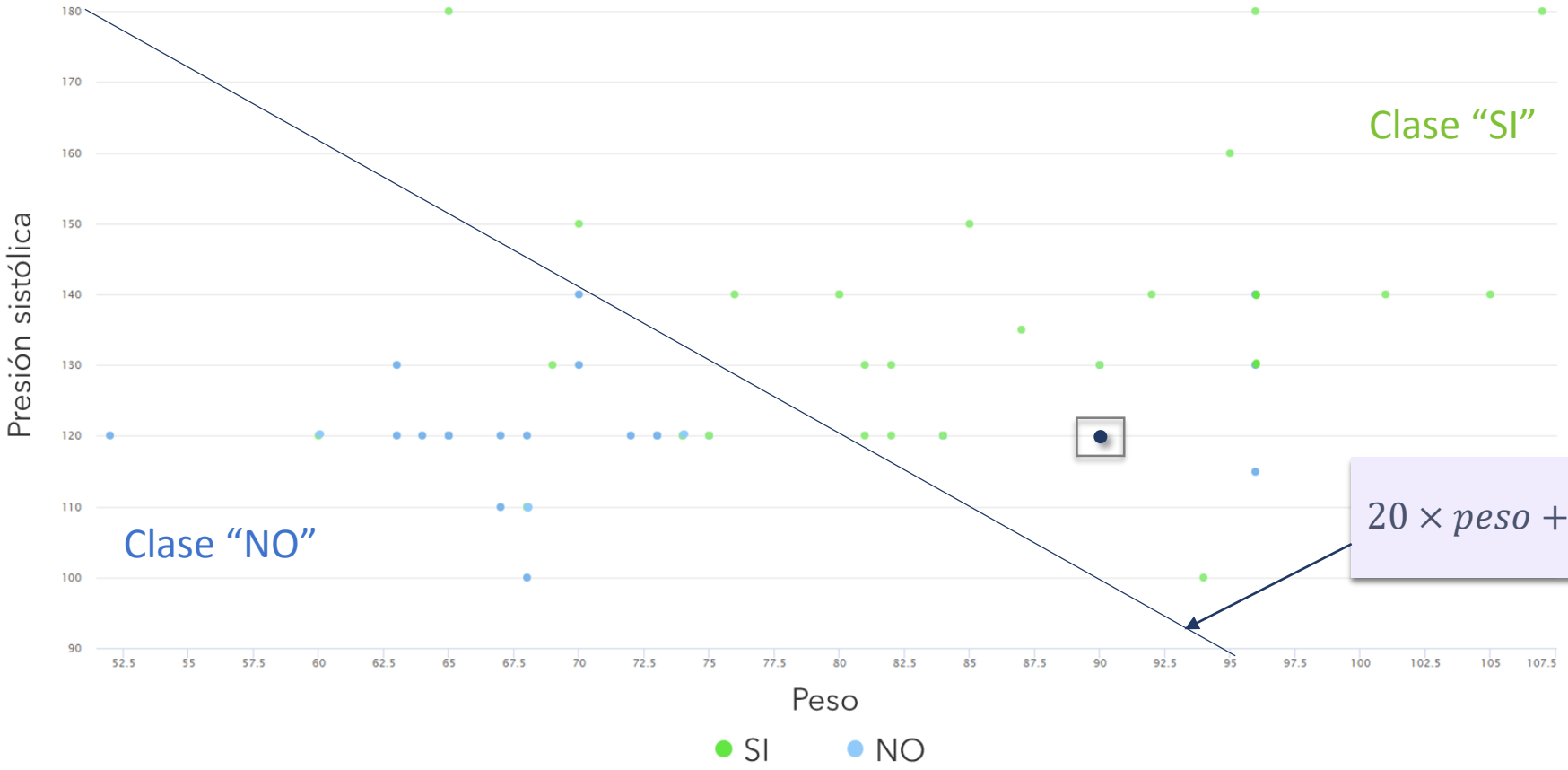
$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

Clase estimada $= f_{umbral}(w_0 + w_1 x_1 + w_2 x_2)$

$$f_{umbral}(h) = \begin{cases} clase\ \text{"SI"} & si\ h \geq 0 \\ clase\ \text{"NO"} & si\ h < 0 \end{cases}$$

← umbral

¿Cómo clasificar?



Discriminante lineal

$$20 \times peso + 10 \times presión\ sistólica - 2800 = 0$$

clase SI $\Rightarrow$ $20 \times peso + 10 \times presión\ sistólica - 2800 \geq 0$

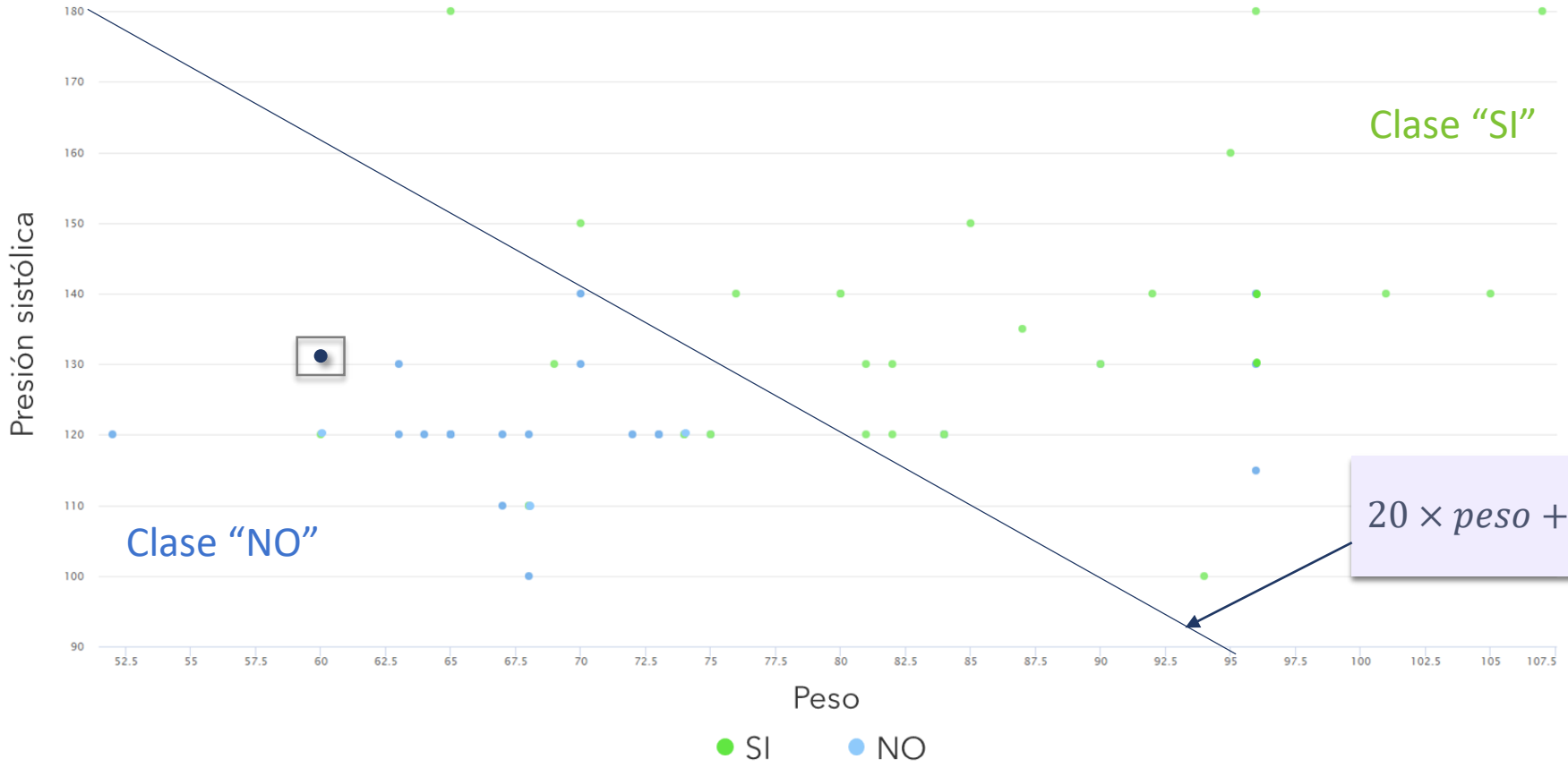
clase NO $\Rightarrow$ $20 \times peso + 10 \times presión\ sistólica - 2800 < 0$

$Si\ (90,120)$:

$20 \times 90 + 10 \times 120 - 2800 = 200$

$\Rightarrow Clase\ SI$

¿Cómo clasificar?



Discriminante lineal

$$20 \times peso + 10 \times presión\ sistólica - 2800 = 0$$

clase SI si: $20 \times peso + 10 \times presión\ sistólica - 2800 \geq 0$

clase NO si: $20 \times peso + 10 \times presión\ sistólica - 2800 < 0$

$Si\ (60,130)$:

$20 \times 60 + 10 \times 130 - 2800 = -300$

$\Rightarrow Clase\ NO$

Algunos algoritmos:

Clasificadores binarios

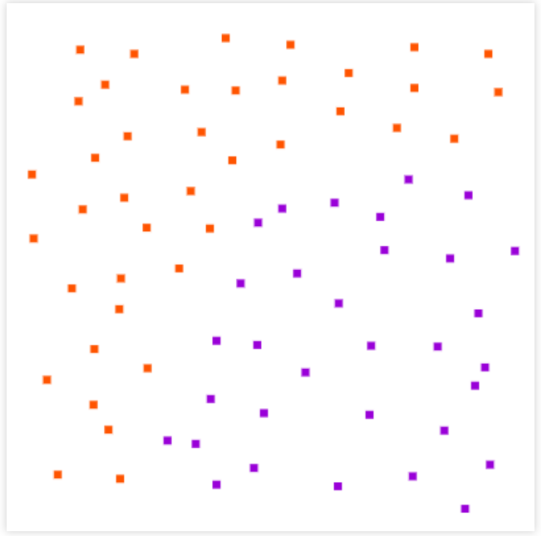
- ❑ Regresión logística
- ❑ Máquinas de vectores de soporte

Clasificadores multiclase

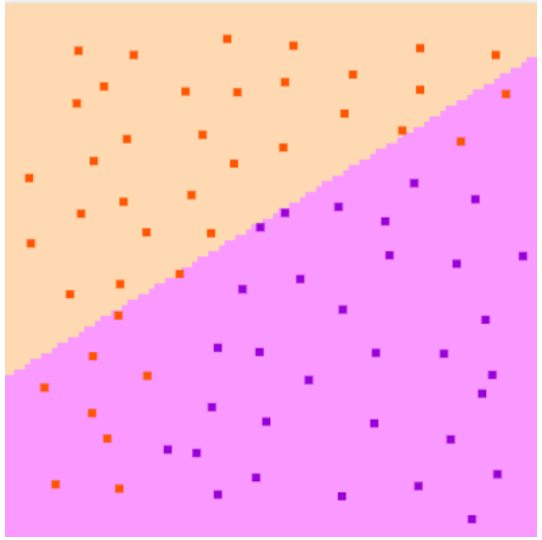
- ❑ Árboles de decisión
- ❑ Redes neuronales
- ❑ Basados en probabilidades (Naïve Bayes)
- ❑ K-vecinos más cercanos
- ❑ Ensembles (metal algoritmos). Ej. Random Forest
- ❑ …

En general…

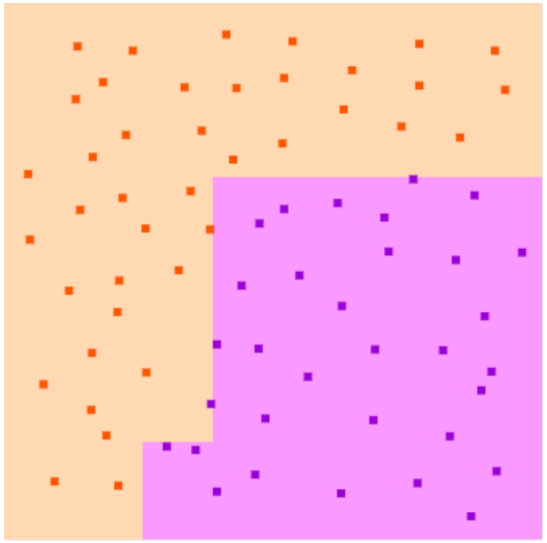
Datos



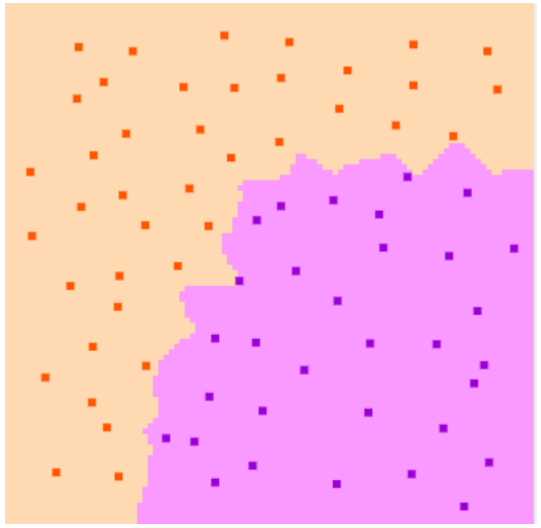
Discriminante lineal



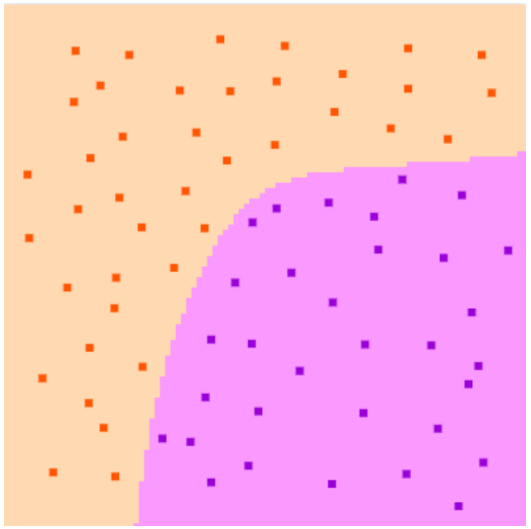
Árbol de decisión



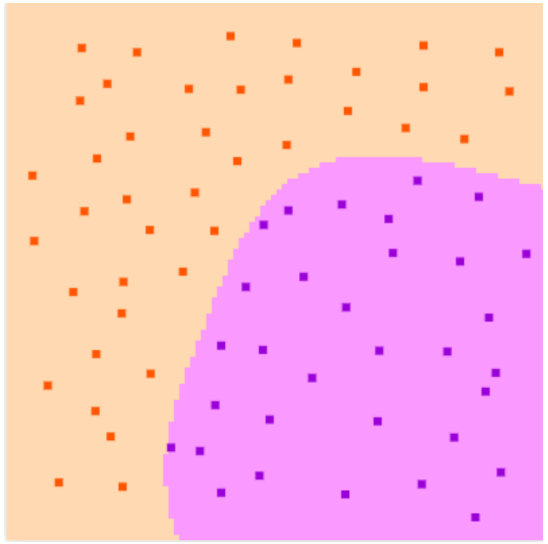
K-vecinos más cercanos



Red neuronal

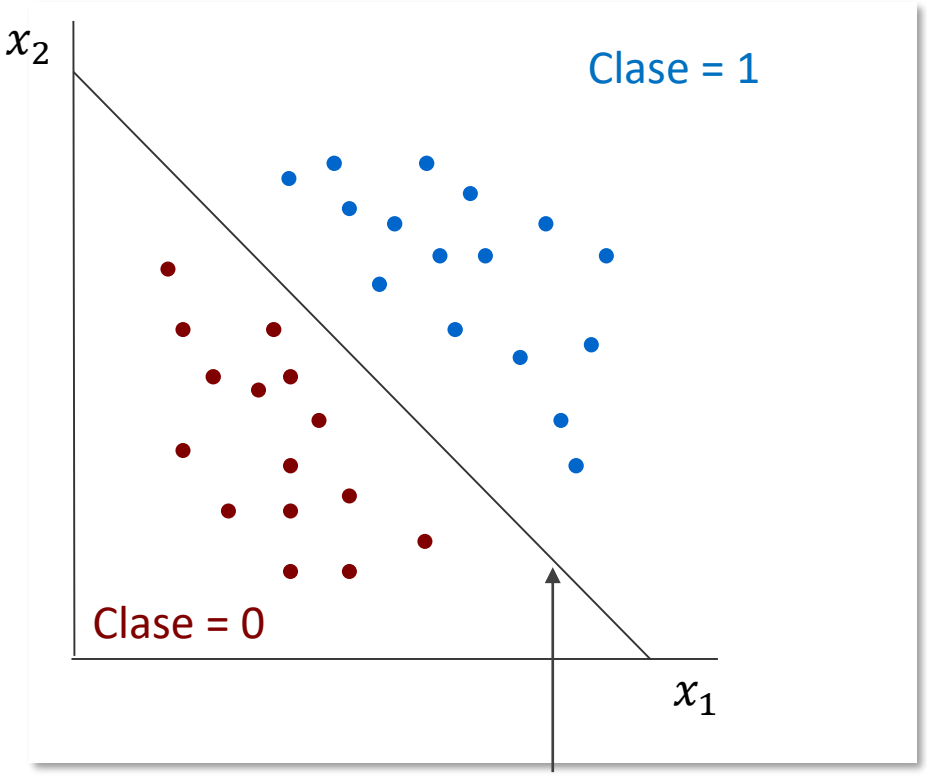


Máquina de vectores de soporte



Machine Learning Playground. https://ml-playground.com/

Problema de clasificación de dos clases:



Límite de decisión lineal

$$w_0 + w_1x_1 + w_2x_2 = 0$$

En vez de una función umbral:

$$f_{umbral}(w_0 + w_1x_1 + w_2x_2)$$

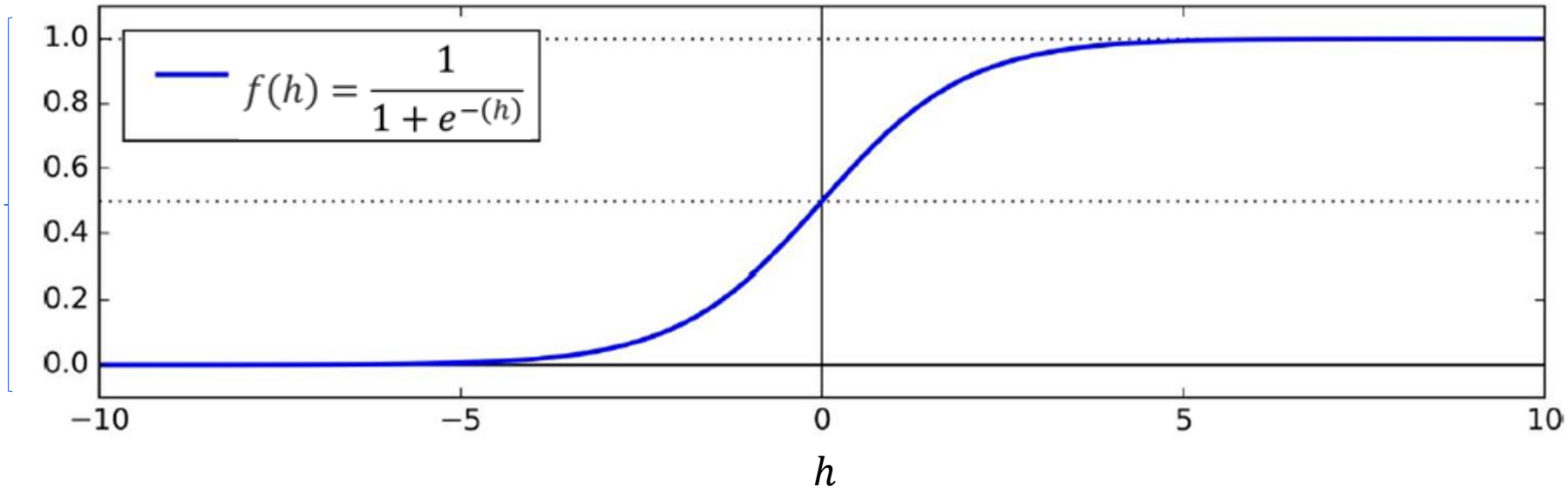
$$f_{umbral}(h) = \begin{cases} clase\ "1" & si\ h \geq 0 \\ clase\ "0" & si\ h < 0 \end{cases}$$

En regresión logística se utiliza la función logística:

$$f_{logística}(w_0 + w_1x_1 + w_2x_2)$$

$$f_{logística}(h) = \frac{1}{1 + e^{-(h)}}$$

Rango = [0,1]

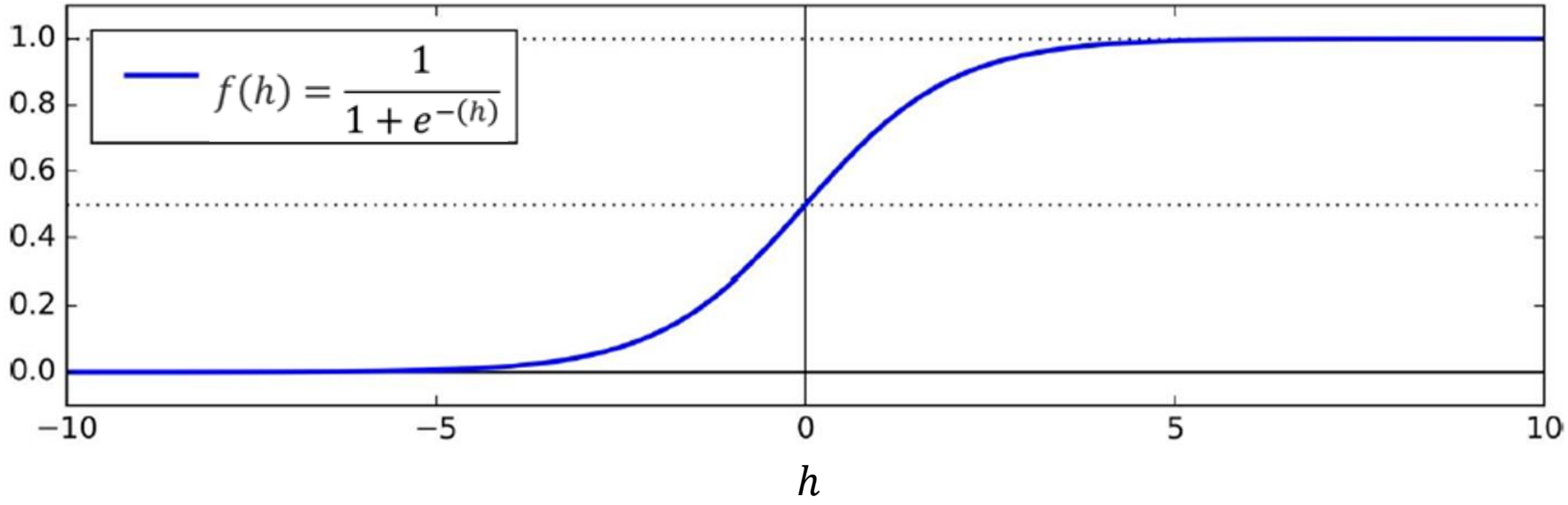


- Puede ser interpretado como la probabilidad de un evento binario.
- Representa la probabilidad de la clase dada la instancia $\boldsymbol{x}$ como:

$$P(y|\boldsymbol{x}) = \frac{1}{1+e^{-(\boldsymbol{w}\cdot\boldsymbol{x}+b)}}) = probabilidad\ estimada\ de\ y = 1, dado\ que\ se\ presenta\ \boldsymbol{x}$$

$$Clase\ 1\colon\ P(y=1|\boldsymbol{x})$$

$$Clase\ 0\colon\ P(y=0|\boldsymbol{x}) = 1 - P(y=1|\boldsymbol{x})$$



$$Límite\ de\ decisión \Rightarrow \quad \boldsymbol{w} \cdot \boldsymbol{x} = 0$$

$$\theta(\boldsymbol{w} \cdot \boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w} \cdot \boldsymbol{x} + b)}} = \frac{1}{1 + e^{0}} = 0.5$$

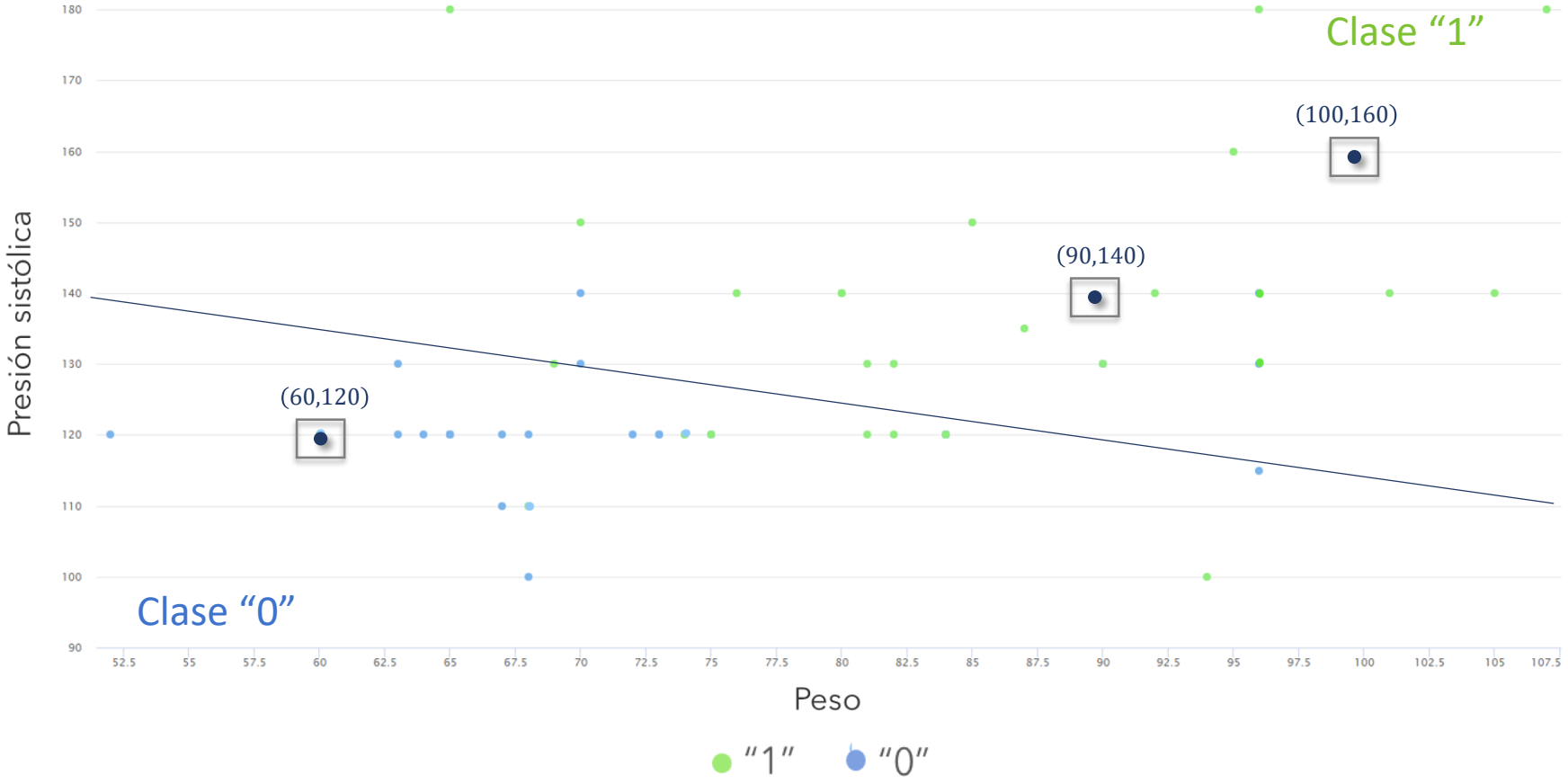
- Para clasificar se utiliza un umbral:

$$predice\ clase\ 1 \quad \text{si} \quad f_{logística}(w_0 + w_1 x_1 + w_2 x_2) \geq 0.5$$

$$predice\ clase\ 0 \quad \text{si} \quad f_{logística}(w_0 + w_1 x_1 + w_2 x_2) < 0.5$$

$$Límite\ de\ decisión = 0.017 \times peso + 0.066 \times presión\ sistólica - 9.15 = 0$$

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-(0.017 \times peso + 0.066 \times presión\ sistólica - 9.15)}}$$



$clase\ 1: \quad f_{logística}(h) \geq 0.5$

$clase\ 0: \quad f_{logística}(h) < 0.5$

$$P(y = 1|\boldsymbol{x} = (90,140)) = \frac{1}{1 + e^{-(0.017\times90\ +\ 0.066\times140\ -\ 9.15)}} = 0.71 \longrightarrow \text{Clase “1”}$$

$$P(y = 1|\boldsymbol{x} = (60,120)) = \frac{1}{1 + e^{-(0.017\times60\ +\ 0.066\times120\ -\ 9.15)}} = 0.40 \longrightarrow \text{Clase “0”}$$

$$P(y = 1|\boldsymbol{x} = (100,160)) = \frac{1}{1 + e^{-(0.017\times100\ +\ 0.066\times160\ -\ 9.15)}} = 0.96 \longrightarrow \text{Clase “1”}$$

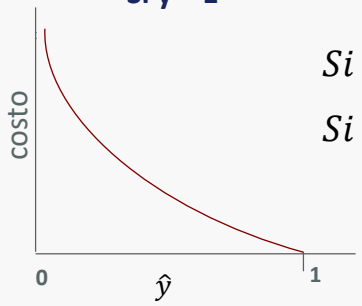
¿Por qué la diferencia en las probabilidades?

¿Cómo determinar los parámetros?

$$Costo\ para\ una\ instancia = \begin{cases} -log(\hat{y}) & si\ y = 1 \\ -log(1\ -\ \hat{y}) & si\ y = 0 \end{cases}$$

Recordar: el $costo(\hat{y}, y)$ mide la discrepancia entre lo que predice el modelo y el valor real.

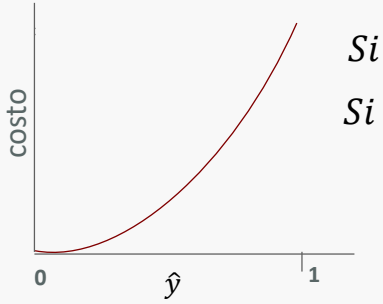
Si y = 1



$Si\ \hat{y} = 1 \Rightarrow\ costo = 0$

$Si\ \hat{y} = 0 \Rightarrow\ costo \longrightarrow \infty$

Si y = 0



$Si\ \hat{y} = 0 \Rightarrow\ costo = 0$

$Si\ \hat{y} = 1 \Rightarrow\ costo \longrightarrow \infty$

Otra representación:

$$Costo(\hat{y}, y) = -ylog(\hat{y}) - (1 - y)log(1\ -\hat{y})$$

Entonces, función de costo de la regresión logística

$$J(\boldsymbol{W}) = -\frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y_i}) +\ (1\ -\ y_i)log(1\ -\hat{y_i})$$

¿Cómo funciona el algoritmo a utilizar?

¿Qué requerimientos tiene?

¿Cómo ajustar sus parámetros?

Aprendizaje supervisado
Tarea de clasificación

Solo atributos numéricos

	x_1	x_2		x_d	y
Instancia 1					
Instancia 2					
Instancia N					

Parámetros del algoritmo:

C, hiperparámetro de regularización

Regresión logística

Procedimiento de Optimización

Función de costo

Costo log (cross entropy binaria)

$$J(\boldsymbol{W}) = -\frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) + C\sum_{j=1}^{d}(w_j)^2$$

Algoritmo de descenso por el gradiente

Entrada = Conjunto de datos $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1..N}$

Comienzo_procedimiento

Inicializar los pesos $\boldsymbol{W}(0)$ *y la constante de aprendizaje* α

Para k = 0 hasta num_iteraciones

Calcular el gradiente $g_k = \nabla J(\boldsymbol{w}(k))$

Actualizar los parámetros:

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) - \alpha g_k$$

Fin-procedimiento

*Salida = {**W**}*

Modelo estimado

$$\hat{y} = función_{logística}(w_0 + w_1 \times x_1 + w_2 \times x_2 + \cdots + w_d \times x_d)$$

$$parámetros\ del\ modelo = \{w_0, w_1, \ldots, w_d\}$$

$$d = No.\ de\ variables$$

Próxima sesión

- ¿Cómo evaluar los modelos?
- Capacidad de generalización, complejidad de modelos y sobreajuste.
- Dilema bias-varianza.
- Regularización. Hiperparámetros.
- Selección de modelos y técnicas de validación.