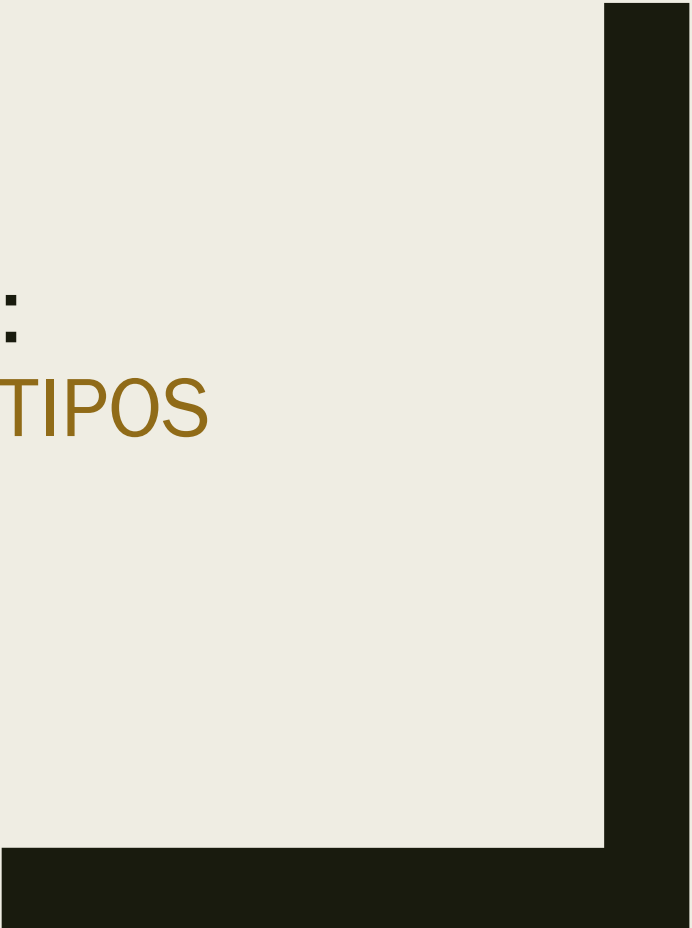




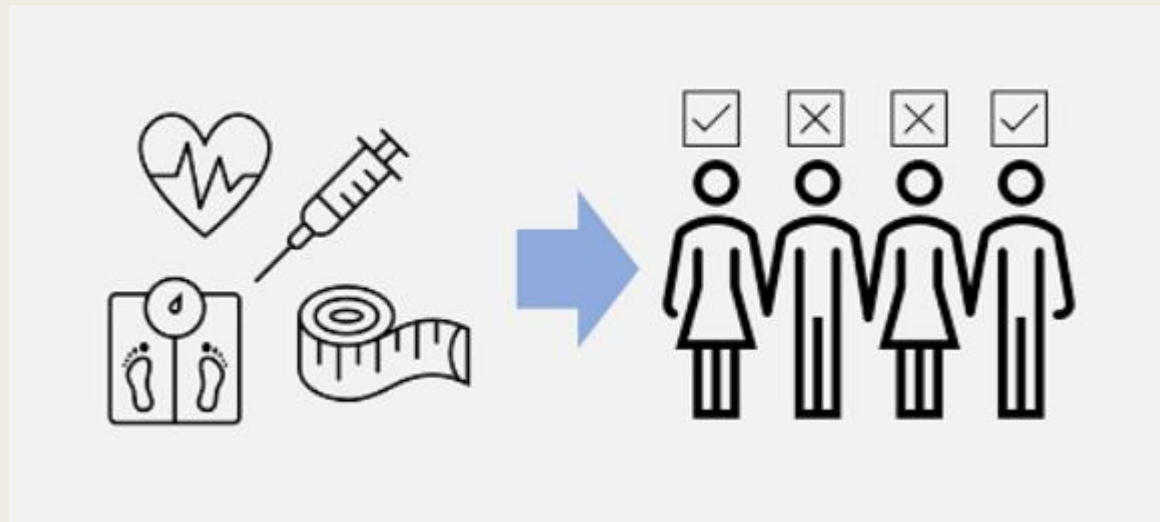
ANALÍTICA AVANZADA DE DATOS: CLASIFICADORES

A. Alejandra Sánchez Manilla
asanchezm.q@gmail.com



Clasificadores

- La clasificación es una forma de aprendizaje automático en la que se entrena un modelo para predecir a qué categoría pertenece un elemento.
- Por ejemplo, una clínica de salud puede utilizar datos de diagnóstico como la altura, el peso, la presión arterial y el nivel de glucosa en sangre de un paciente para predecir si es diabético o no.



Clasificadores

- **Clasificación:** Determinar la clase o grupo al cual pertenece una cosa
- **Un clasificador sencillo:** permite clasificar automáticamente naranjas y limones.



- ¿Qué se necesita?
Un criterio de decisión:
 - Tamaño
 - Color
 - Textura

Clasificadores

- Con medir el diámetro, es posible decidir si es una naranja o un limón.



$$x^1 = 3.9$$



$$x^2 = 4.1$$



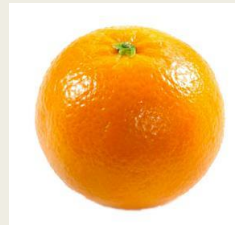
$$x^3 = 4.4$$



$$x^4 = 4.7$$



$$x^5 = 4.5$$



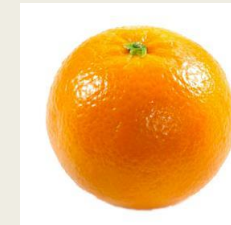
$$x^6 = 6.1$$



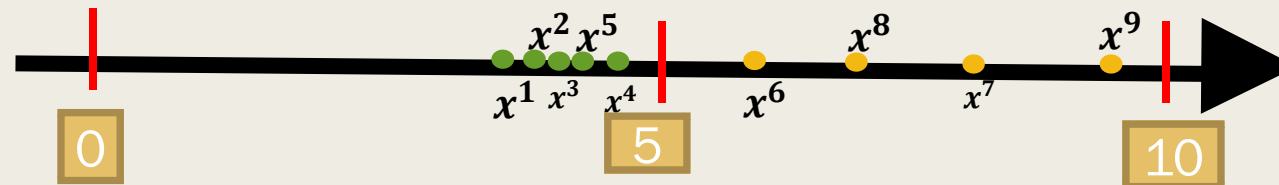
$$x^7 = 8.5$$



$$x^8 = 7.1$$



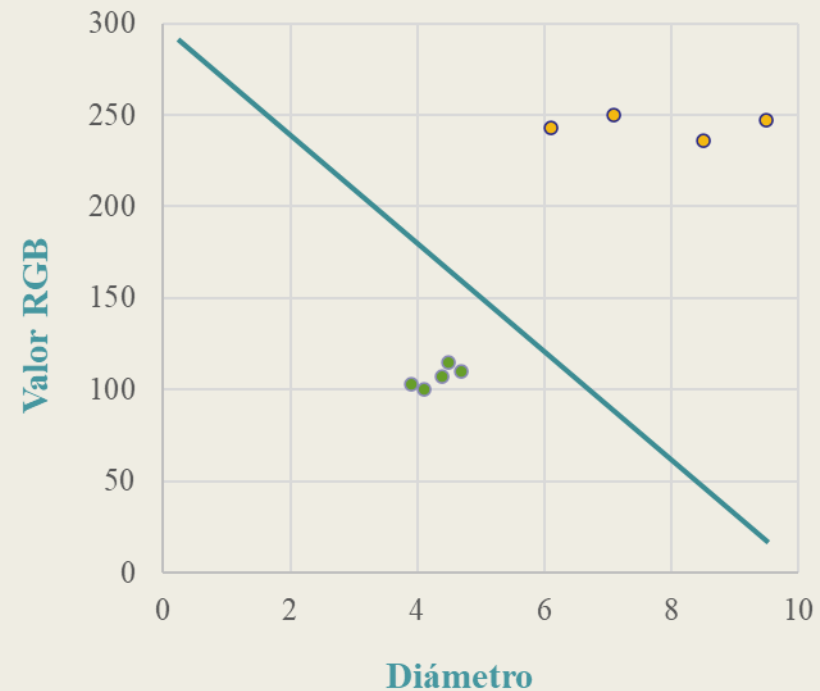
$$x^9 = 9.5$$



Clasificadores

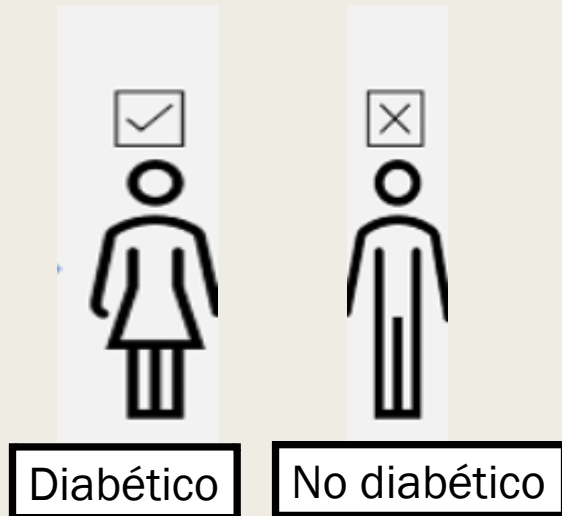
- Con más información del objeto, la clasificación puede mejorar

$x^1 = \begin{bmatrix} 3.9 \\ (103,158,42) \end{bmatrix}$	$x^6 = \begin{bmatrix} 6.1 \\ (243,184,17) \end{bmatrix}$
$x^2 = \begin{bmatrix} 4.1 \\ (100,151,45) \end{bmatrix}$	$x^7 = \begin{bmatrix} 8.5 \\ (236,190,23) \end{bmatrix}$
$x^3 = \begin{bmatrix} 4.4 \\ (107,160,40) \end{bmatrix}$	$x^8 = \begin{bmatrix} 7.1 \\ (250,180,12) \end{bmatrix}$
$x^4 = \begin{bmatrix} 4.7 \\ (110,150,48) \end{bmatrix}$	$x^9 = \begin{bmatrix} 9.5 \\ (247,188,22) \end{bmatrix}$
$x^5 = \begin{bmatrix} 4.5 \\ (115,151,38) \end{bmatrix}$	



Clasificadores

- La clasificación binaria es una clasificación con dos categorías.




La predicción de clase se realiza determinando la probabilidad para cada clase posible como un valor entre 0 y 1

La probabilidad total para todas las clases es 1

Si la probabilidad que un paciente sea diabético es **0.3**, entonces existe una probabilidad de **0.7** de que el paciente no sea diabético.

Clasificadores





ENTRENAMIENTO Y EVALUACIÓN



Clasificadores

Entrenamiento y evaluación de un modelo de clasificación

- La clasificación es un ejemplo de técnica de *aprendizaje automático supervisado*, se basa en datos que incluyen:
 - Valores de características conocidos (por ejemplo, mediciones diagnósticas de pacientes)
 - Valores de etiquetas conocidos (por ejemplo, una clasificación de no diabético o diabético)
- Se utiliza un algoritmo de clasificación para ajustar un subconjunto de datos a una función que puede calcular la probabilidad de cada etiqueta de clase a partir de los valores de las características.

Entrenamiento y evaluación

Entrenamiento y evaluación de un modelo de clasificación

- Los datos restantes se utilizan para evaluar el modelo comparando las predicciones que genera a partir de las características con las etiquetas de clase conocidas.

Por ejemplo:

- Supongamos que tenemos los siguientes datos de un paciente, que consisten en una única característica (nivel de glucosa en sangre) y una etiqueta de clase 0 para no diabético, 1 para diabético.

nivel de glucosa en sangre

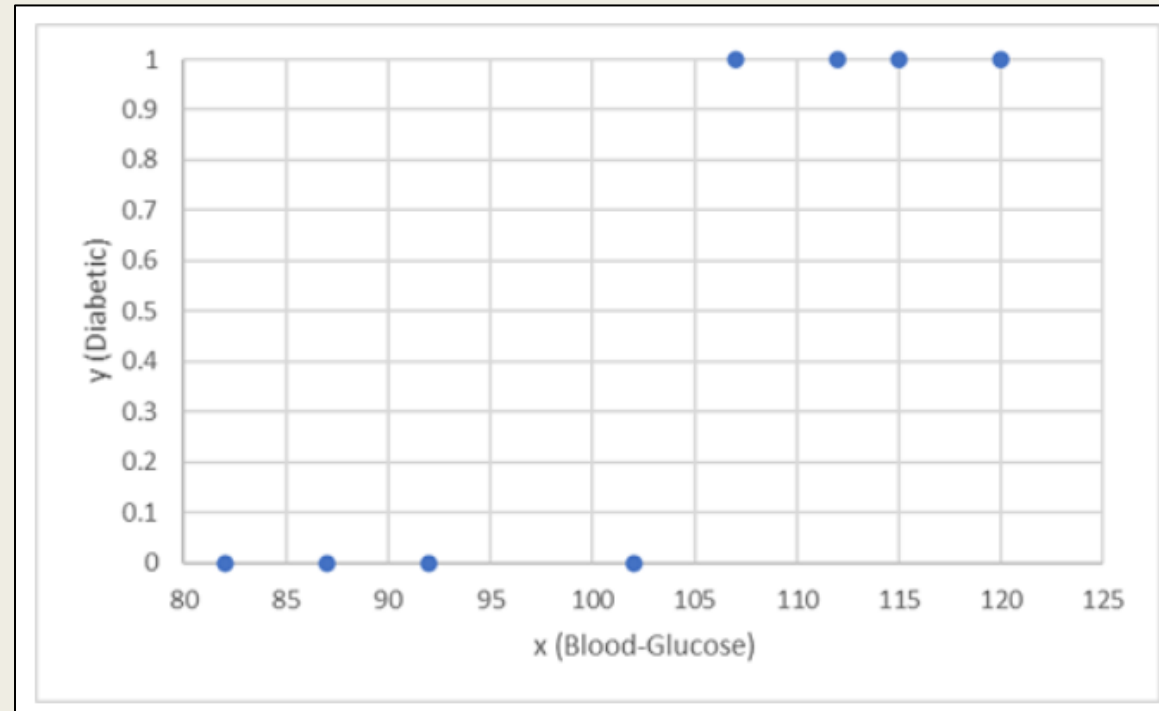
clase 0

clase 1

Entrenamiento y evaluación

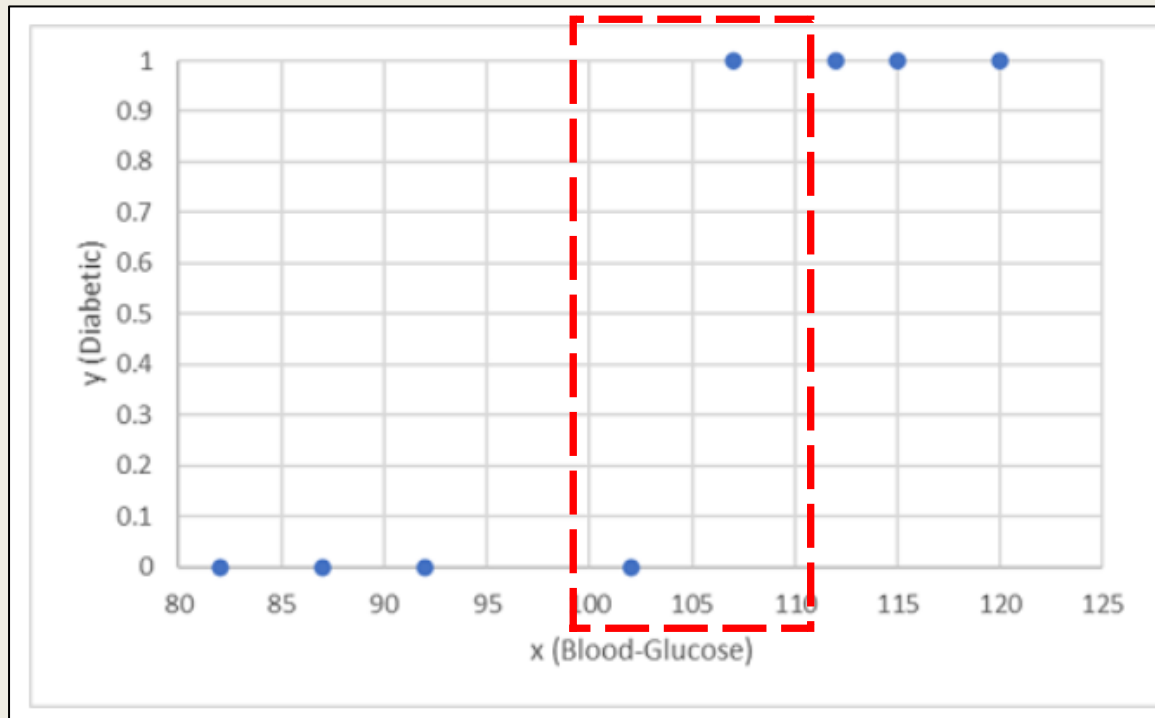
Blood-Glucose	Diabetic
82	0
92	0
112	1
102	0
115	1
107	1
87	0
120	1
83	0
119	1
104	1
105	0
86	0
109	1

Utilizaremos las ocho primeras observaciones para entrenar un modelo de clasificación, y empezaremos trazando la característica de glucosa en sangre (que llamaremos x) y la etiqueta de diabético predicha (que llamaremos y).



Entrenamiento y evaluación

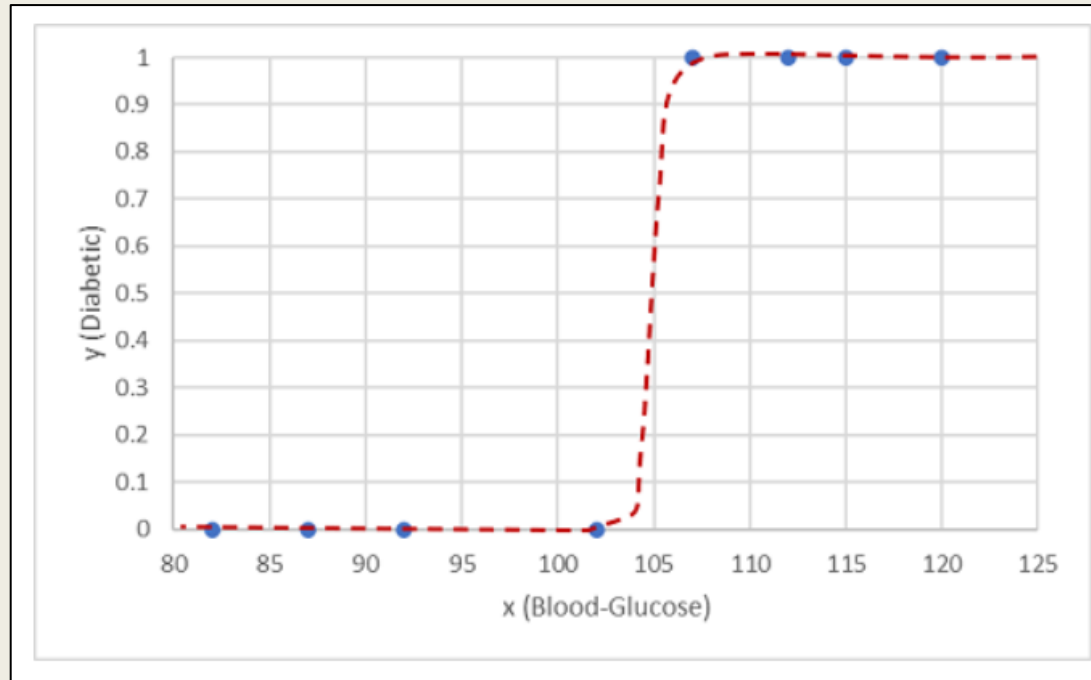
- Lo que necesitamos es una **función que calcule un valor de probabilidad para y basado en x** (en otras palabras, necesitamos la función $f(x) = y$).
- En la gráfica se ve que los pacientes con un *nivel bajo* de glucosa en sangre *no son diabéticos*, mientras que los pacientes con un *nivel alto* de glucosa en sangre son *diabéticos*.



Parece que cuanto más alto es el nivel de glucosa en sangre, más probable es que un paciente sea diabético, con **el punto de inflexión** en algún lugar entre 100 y 110.

Entrenamiento y evaluación

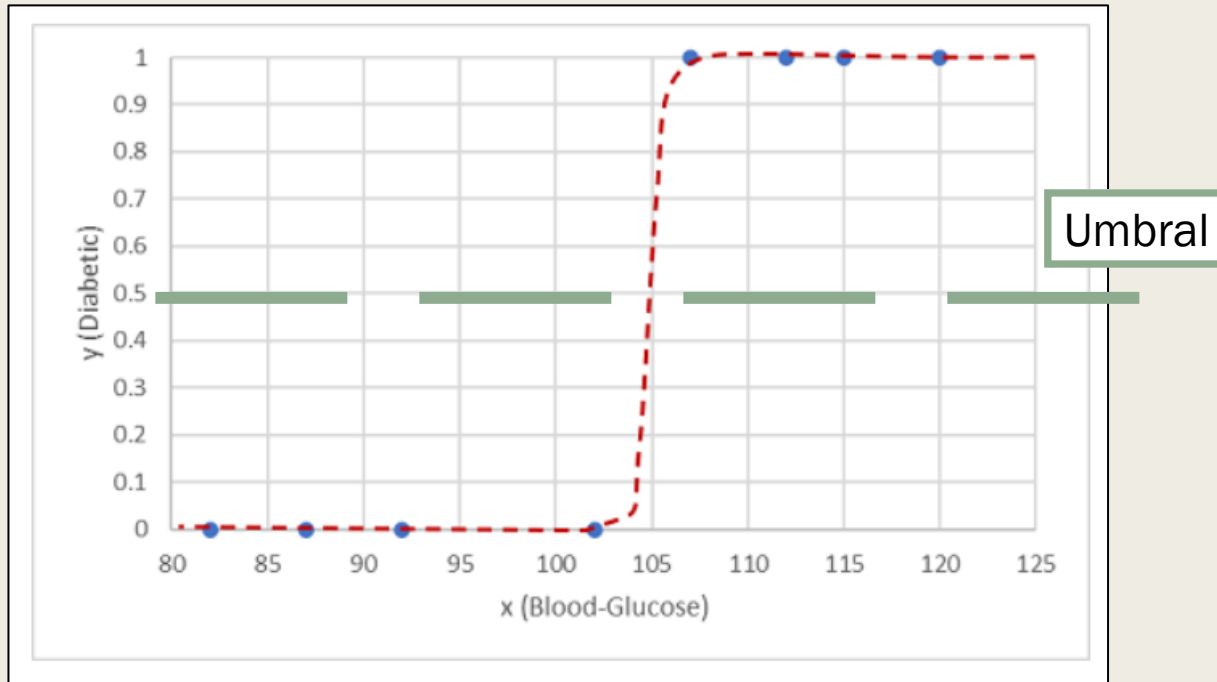
- Necesitamos ajustar a estos valores una función que calcule un valor entre 0 y 1 para y
- Una de estas funciones es la función logística, que forma una curva sigmoideal (en forma de S), como ésta:



Podemos utilizar la función:

- Para calcular un valor de probabilidad de que y sea positivo, lo que significa que el paciente es diabético a partir de cualquier valor de x

Entrenamiento y evaluación



Podemos utilizar la función:

- Establecer un valor umbral de 0.5 como punto de corte para la predicción de la etiqueta de clase
- Los puntos trazados por debajo de la línea de umbral arrojarán una predicción de clase de 0 -no diabético- y los puntos por encima de la línea se predecirán como 1 -diabético-.

Entrenamiento y evaluación

Ahora podemos comparar las predicciones de etiquetas basadas en la función logística encapsulada en el modelo (que llamaremos \hat{y} , o "y-hat") con las etiquetas de clase reales (y).

x	y	\hat{y}
83	0	0
119	1	1
104	1	0
105	0	1
86	0	0
109	1	1

Entrenamiento y evaluación

Evaluar los modelos de clasificación

- Después de haber entrenado un modelo de clasificación, debemos **evaluar su rendimiento** con un conjunto de datos nuevos y desconocidos.
- Retomando el ejemplo de los pacientes (diabético, no diabético) en función de su nivel de glucosa en la sangre. Ahora, al aplicarlo a unos datos que no forman parte del conjunto de entrenamiento, obtendremos las siguientes predicciones:

Entrenamiento y evaluación

Glucosa en sangre	Diabético?	Predicción del modelo
x	y	\hat{y}
83	0	0
119	1	1
104	1	0
105	0	1
86	0	0
109	1	1

- Calcular simplemente cuántas predicciones fueron correctas es a veces engañoso o demasiado simplista para que comprendamos el tipo de errores que cometerá en el mundo real.

Entrenamiento y evaluación

- Para obtener información más detallada, podemos tabular los resultados en una estructura llamada matriz de confusión:

		Predicted	
		0	1
Actual	0	2	1
	1	1	2

La matriz de confusión muestra el número total de casos en los que:

- El modelo predijo 0 y la etiqueta real es 0 (**verdaderos negativos**; *arriba a la izquierda*)
- El modelo predijo 1 y la etiqueta real es 1 (**verdaderos positivos**; *abajo a la derecha*)
- El modelo predijo 0 y la etiqueta real es 1 (**falsos negativos**; *abajo a la izquierda*)
- El modelo predijo 1 y la etiqueta real es 0 (**falsos positivos**; *arriba a la derecha*)

Entrenamiento y evaluación


		Predicted	
		0	1
Actual	0	2	1
	1	1	2

Las celdas de la matriz de confusión suelen estar sombreadas, de modo que *los valores más altos tienen un sombreado más intenso*.


Así es más fácil ver una fuerte tendencia diagonal de arriba a la izquierda a abajo a la derecha, destacando las celdas en las que el valor predicho y el valor real son iguales.

A partir de estos valores básicos, puede calcular una serie de métricas que le ayudarán a evaluar el rendimiento del modelo. Por ejemplo:

- **Accuracy:** $(TP+TN)/(TP+TN+FP+FN)$ - De todas las predicciones, ¿cuántas fueron correctas?
- **Recall:** $TP/(TP+FN)$ - de todos los casos positivos, ¿cuántos identificó el modelo?
- **Precision:** $TP/(TP+FP)$ - De todos los casos que el modelo predijo que serían positivos, ¿cuántos son realmente positivos?



MODELOS DE CLASIFICACIÓN MULTICLASE



Clasificación multiclase

Crear modelos de clasificación multiclase

- También es posible crear modelos de clasificación multiclase, en los que hay más de dos clases posibles
- Por ejemplo, la clínica de salud podría ampliar el modelo de diabetes para clasificar a los pacientes como:
 - *No diabéticos*
 - *Diabético de tipo 1*
 - *Diabético de tipo 2*
- Los valores de probabilidad de cada clase seguirían sumando un total de 1, ya que el paciente sólo está en una de las tres clases, y el modelo predeciría la clase más probable

Clasificación multiclase

Utilización de modelos de clasificación multiclase

- La clasificación multiclase puede considerarse como una combinación de varios clasificadores binarios. Hay dos formas de abordar el problema:

Uno frente al resto (One vs Rest, OVR)

Se crea un clasificador para cada posible valor de clase, con un resultado positivo para los casos en los que la predicción es esta clase, y predicciones negativas para los casos en los que la predicción es cualquier otra clase.

Por ejemplo, un problema de clasificación con cuatro posibles clases de forma (cuadrado, círculo, triángulo, hexágono) requeriría cuatro clasificadores que predijeran

Clasificación multiclase

Uno contra uno (One vs One, OVO)

Se crea un clasificador para cada posible par de clases.

El problema de clasificación con cuatro clases de formas requeriría los siguientes clasificadores binarios:

- *cuadrado o círculo*
- *cuadrado o triángulo*
- *cuadrado o hexágono*
- *círculo o triángulo*
- *círculo o hexágono*
- *triángulo o hexágono*

Clasificación multiclase

- En ambos enfoques, el modelo global debe tener en cuenta todas estas predicciones para determinar a qué categoría pertenece el elemento
- En la mayoría de los marcos de aprendizaje automático, incluido *scikit-learn*, la implementación de un modelo de clasificación multiclase no es significativamente más compleja que la clasificación binaria
- En la mayoría de los casos, los estimadores utilizados para la clasificación binaria admiten implícitamente la clasificación multiclase abstrayendo un algoritmo OVR, un algoritmo OVO o permitiendo la elección de cualquiera de ellos.