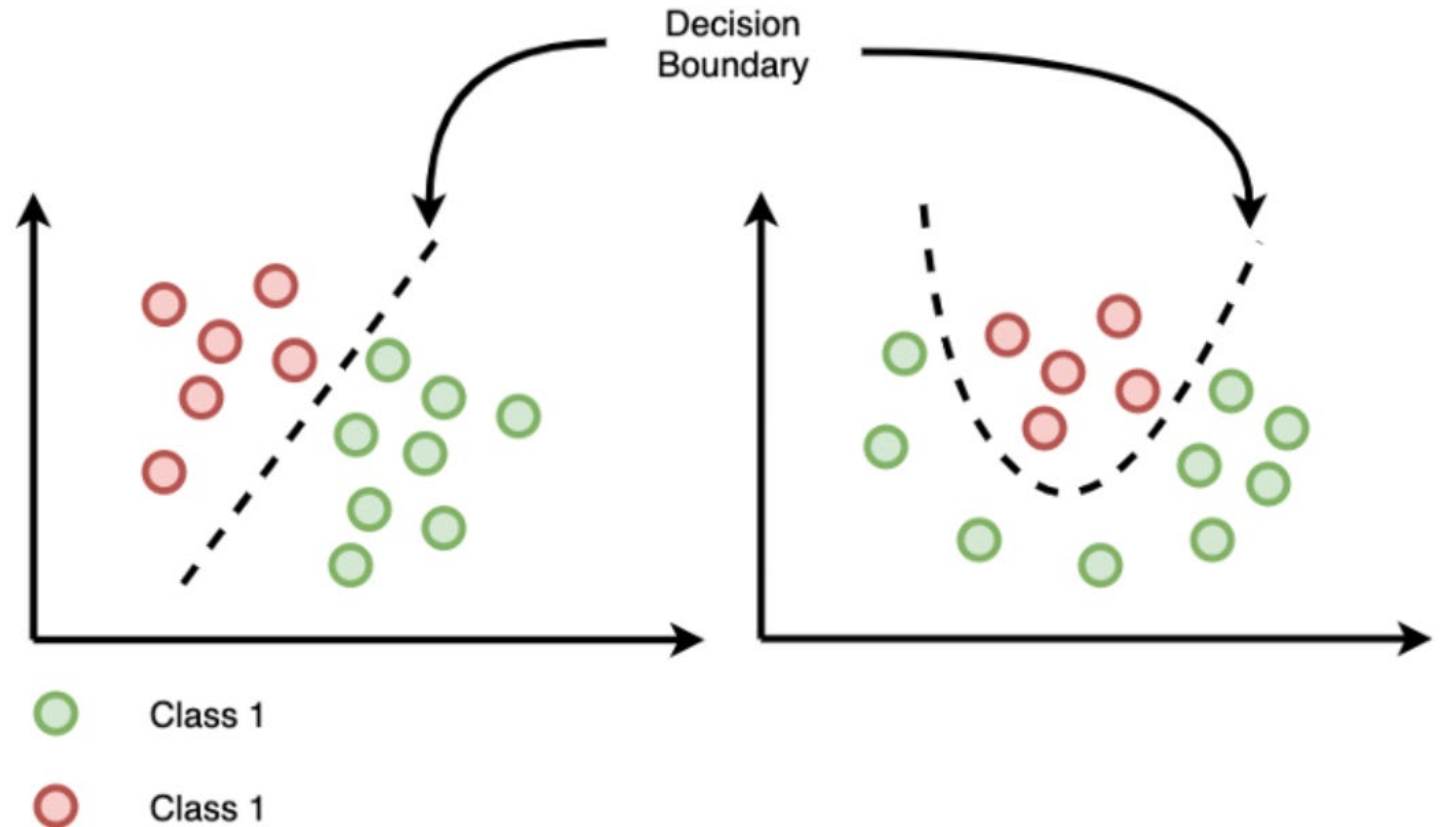


Clasificación lineal: Regresión Logística

Claudia Chávez O.

Limites de decisión (decision boundaries)

- Un modelo de clasificación se encarga de identificar a qué clase pertenece cada registro.



Limite recto

- Clasificadores lineales
- Regresión, logística, svm

Limite no recto

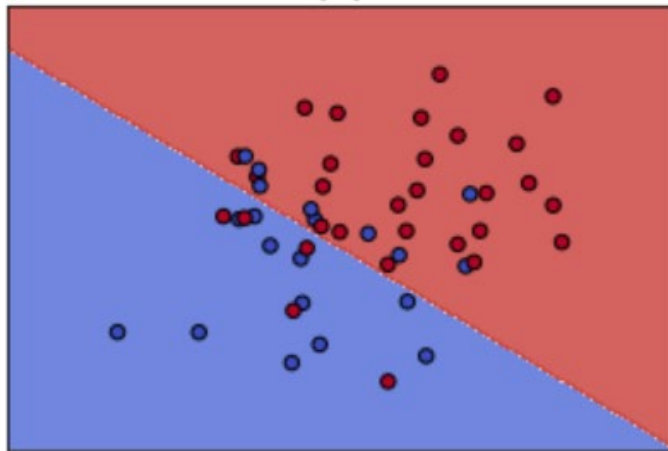
- Clasificadores no lineales
- Árboles de decisión y redes neuronales

Clasificadores

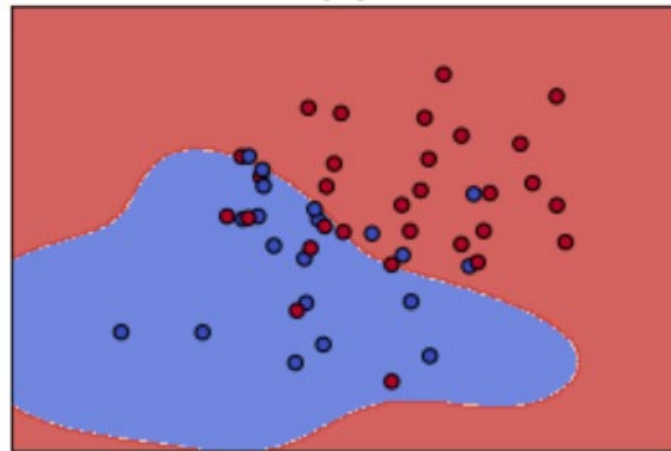


¿Cuál(es)
es/son
lineal/es?

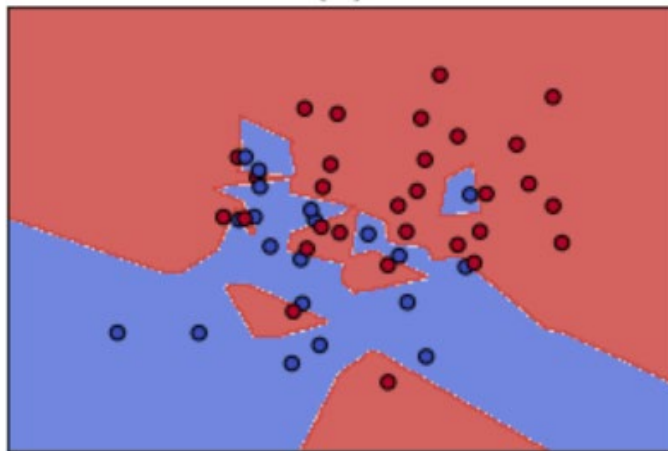
(1)



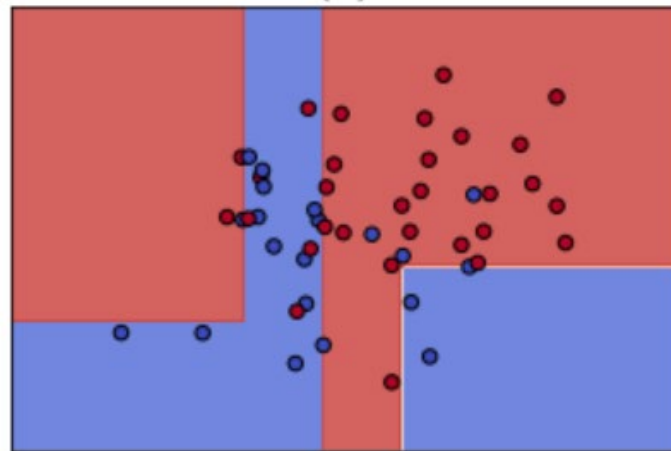
(2)




(3)



(4)



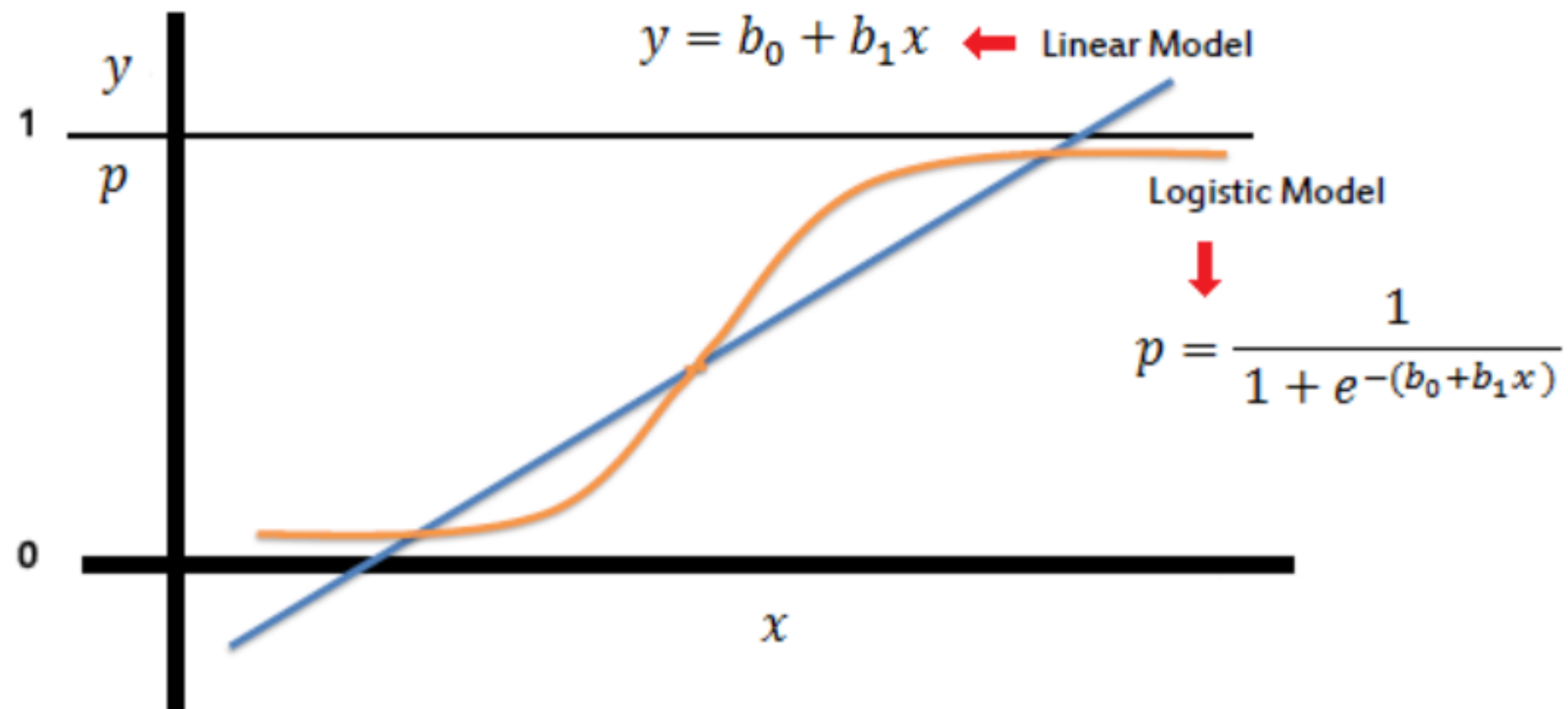


Regresión logística

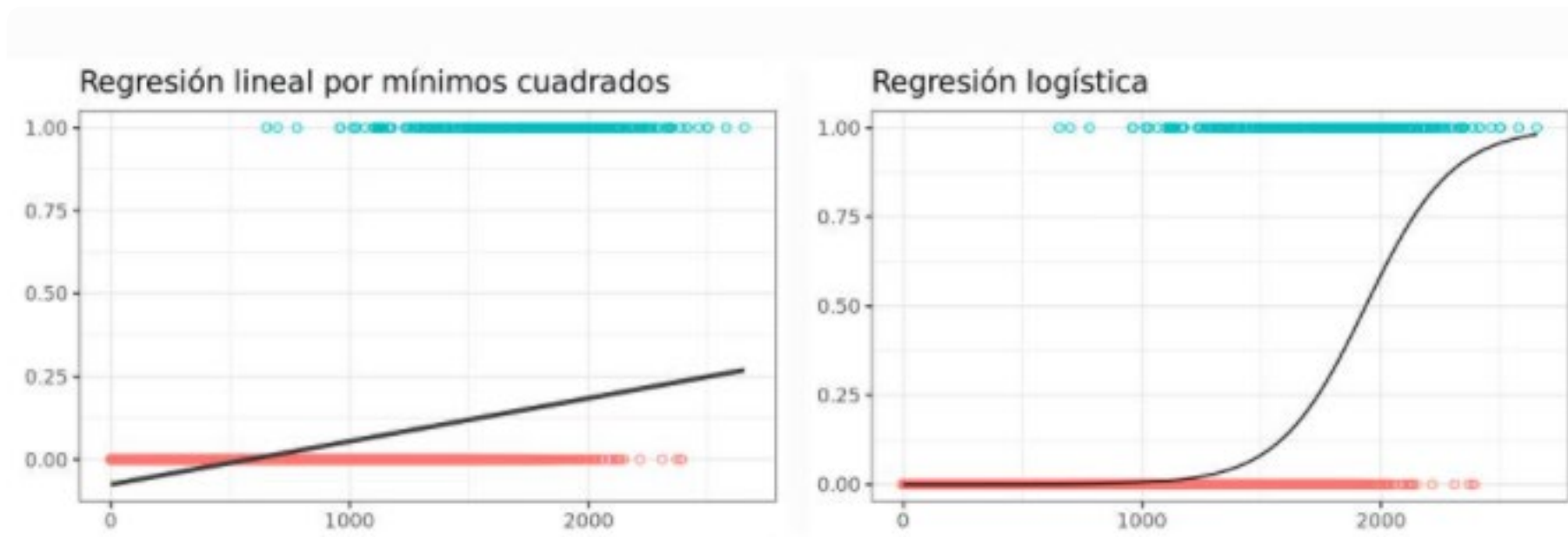
- Ya vieron regresión en el modulo anterior, cual es la diferencia ?
- Requerimos un valor acotado de la variable de respuesta, por lo que parece razonable pasarlo a una probabilidad de ocurrencia para esto la regresión logística.
- Sigue el mismo procedimiento de la regresión lineal múltiple, solo que la variable de respuesta no es continua, sino de tipo dicotómica.

Problemas de la respuesta binaria en regresión lineal múltiple

- Violación de supuestos de la regresión lineal:
 - La distribución de los errores estimados no es normal.
 - No se puede hacer una correcta interpretación de los coeficientes resultantes en términos de probabilidad, en la regresión logística sí.



Modelo de regression lineal vs regression logística para una variable binaria. La línea gris representa la recta de regresión (el modelo)



Regresión logística

- La Regresión Logística es un método estadístico para predecir clases binarias. El resultado o variable objetivo es de naturaleza dicotómica. Dicotómica significa que solo hay dos clases posibles.
- La Regresión Logística es uno de los algoritmos de Machine Learning más simples y más utilizados para la clasificación de dos clases.
- Es fácil de implementar y se puede usar como línea de base para cualquier problema de clasificación binaria.

Regresión logística, ¿para que sirve?

- Teniendo una variable dicotómica como variable de respuesta (éxito /fracaso), y se desea evaluar el efecto de otras variables independientes sobre ella, la regresión logística binaria sirve para:
 - Estimar la probabilidad de que se presente el evento de interés (por ejemplo, tener éxito en el primer semestre), dado los valores de las variables independientes,
 - Evaluar la influencia que cada variable independiente tiene sobre la respuesta en forma de OR (ODD RATIO). Una OR mayor que uno indica aumento en la probabilidad del evento y una OR menor que uno, implica una disminución.

Tipos de regresión logística

- Regresión Logística **Binaria**: la variable objetivo tiene solo dos resultados posibles, Lluvia o NO Lluvia, Sube o Baja.
- Regresión Logística Multinomial: la variable objetivo tiene tres o más categorías **nominales**, como predecir el tipo de vino.
- Regresión Logística Ordinal: la variable objetivo tiene tres o más categorías **ordinales**, como clasificar un restaurante o un producto del 1 al 5.

Ejemplos en los que se ha aplicado regresión logística

- **Credit Scoring**
- Modelos utilizados para la calificación crediticia. En estos modelos es común utilizar técnicas de reducción de variables.
- Este problema es de fácil solución en regresión logística, ya que se puede explicar la influencia de las variables.
- Hasta el día de hoy es de mucha utilidad en este ámbito por sobre los modelos más complejos.

Medicina

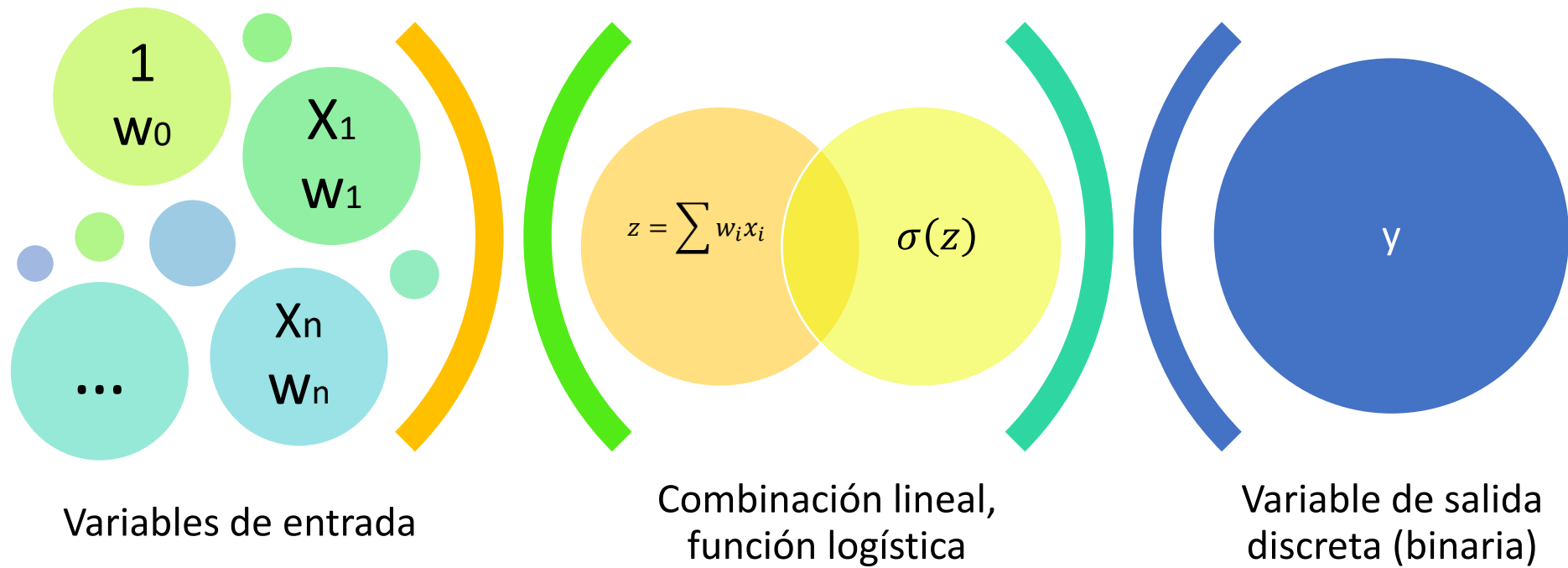
- Miroculus, es una empresa que desarrolla kits de análisis de sangre express, firmó un acuerdo con Microsoft para desarrollar un algoritmo **que identifique la relación entre ciertos micro-ARN y genes en documentos académicos.**
- Para esto utilizaron una base de datos de artículos científicos y aplicaron métodos de análisis de texto para obtener vectores de características.
- Se consideraron como modelos algoritmos como la regresión logística, la máquina de vectores de soporte y el bosque aleatorio. Se seleccionó la regresión logística porque demostró los mejores resultados en velocidad y precisión.



Hotel booking

- Booking.com es una de las plataformas más relevantes de reserva en línea de hoteles.
 - Posee algoritmos de aprendizaje en toda la pagina web.
 - Partieron con algoritmos simples, por ejemplo regresión logística.
- Por ejemplo, todos los datos que tienen son de dónde es el usuario y hacia dónde quiere ir. La regresión logística es ideal para tales necesidades.

Estructura de una regresión logística



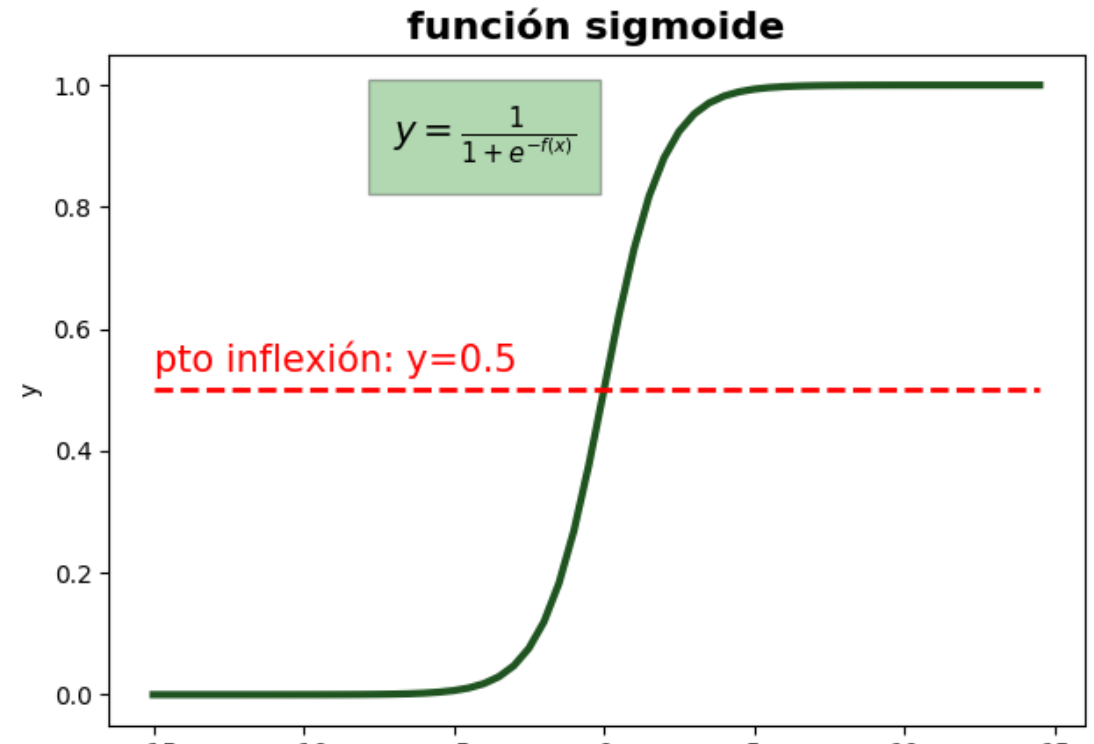
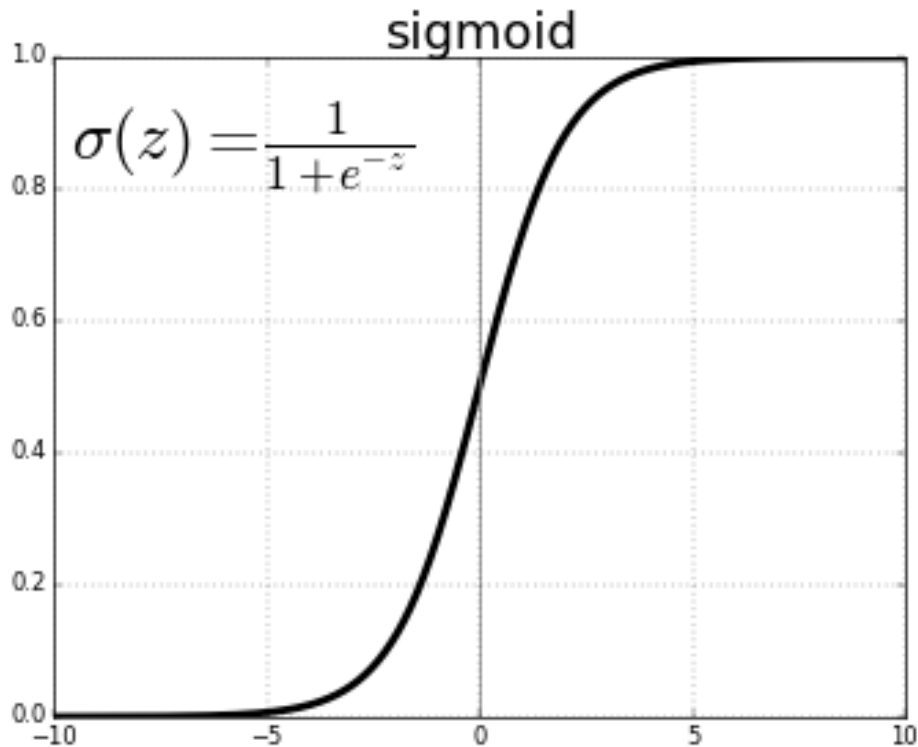
Formalmente esto es ...

$$y = \sigma(z) = \sigma(WX) = \sigma\left(\sum (w_i x_i)\right) = \sigma\left(\sum (w_0 x_0 + w_1 x_1 + \dots + w_n x_n)\right)$$

, donde y es la variable binaria que queremos estimar, σ es la función logística, w son los coeficientes de la combinación lineal (pesos), y x corresponde a las variables de entrada del modelo/algoritmo.

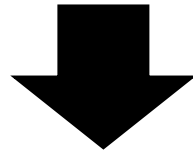
Función logística

También se denomina sigmoide



Volviendo a lo anterior

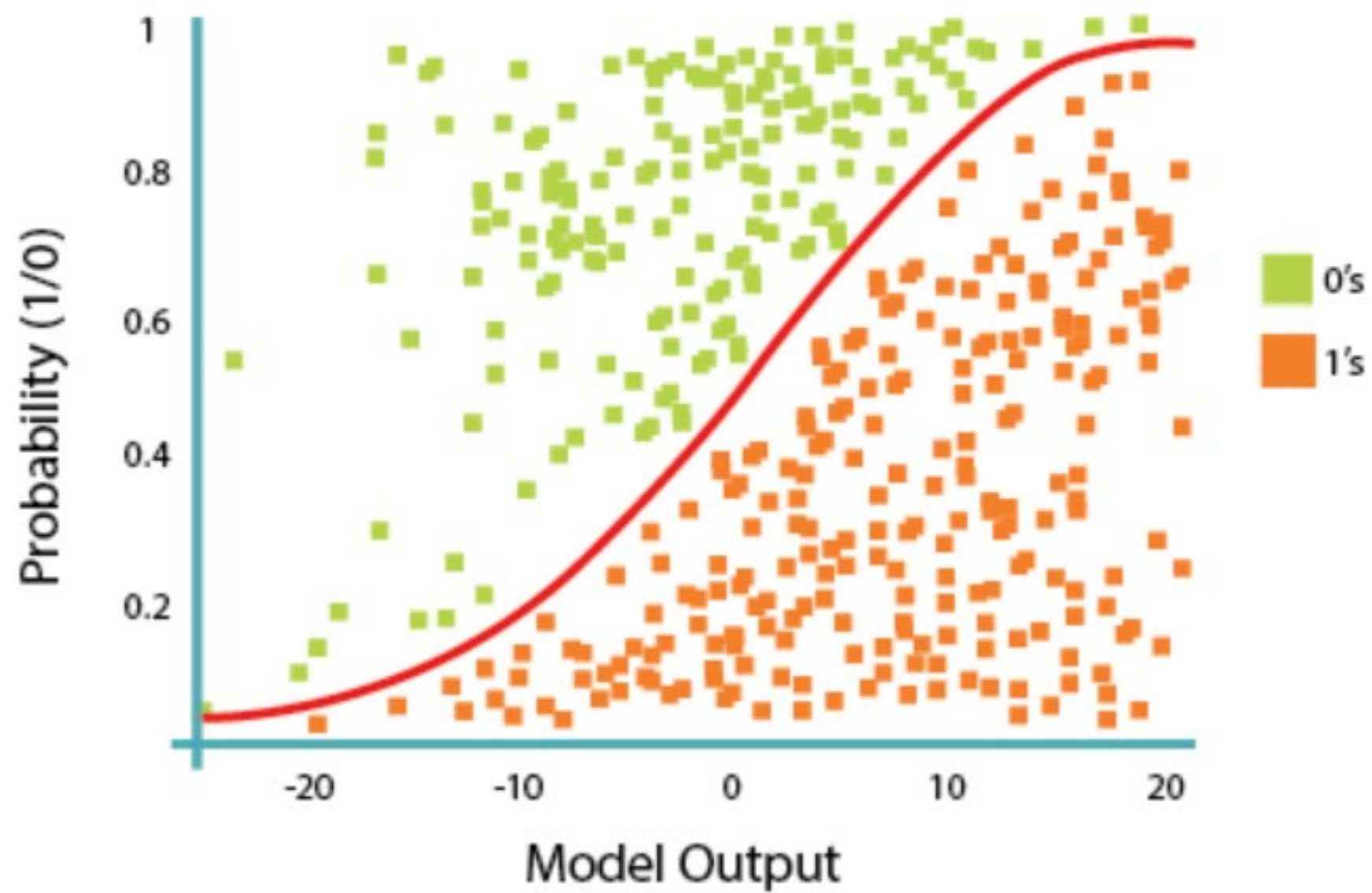
$$y = \sigma(z) = \sigma(WX) = \sigma\left(\sum (w_i x_i)\right) = \sigma\left(\sum (w_0 x_0 + w_1 x_1 + \dots + w_n x_n)\right)$$



$$y = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + \dots + w_n x_n)}}$$

¿Qué
problema
de
clasificación
queremos
resolver?

- Ejemplos: deserción estudiantil, pago o no de un crédito, identificación de una imagen, etc.
- Así, el problema de clasificación binaria es identificar la probabilidad (p) de que un evento ocurra o en otras palabras la probabilidad de que una observación (individuo, etc.) pertenezca a una clase particular.



Regresión logística

- Definamos que queremos estimar la probabilidad de que un evento este presente, esto es

$$p=P(Y=1)$$

- Si lo quisiéramos hacer con regresión lineal, tendríamos:

$$p=\alpha+\beta x$$

- Los resultados se encontrarán en rangos negativos o mayores que 1, para solucionar esto y acotar el resultado a 0,1 (probabilidad) se define una función llamado logito.

Regresión Logística

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

,donde p es la probabilidad de que ocurra el evento de. Dado el valor de las variables independientes, podemos calcular directamente la estimación de la probabilidad de que ocurra el evento de interés de la siguiente forma:

$$\hat{p} = \frac{e^{suma}}{1 + e^{suma}}; \text{ donde } suma = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \dots + \hat{b}_kx_k$$

Regresión Logística

The diagram illustrates the logistic regression equation: $\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$. It includes labels for the logit function, the probability of the event of interest, the parameters, and the independent variables.

logit → $\text{logit}(p)$

Probabilidad de evento de interés → p

Parámetros → $\beta_0, \beta_1, \beta_2, \dots, \beta_n$

Variables independientes → x_1, x_2, \dots, x_n

$$\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Estimación de parámetros

- Nuestro problema al igual que en la regresión lineal es la estimación de los parámetros de la recta para el logit definido como respuesta.
- Método de Máxima Verosimilitud
 - El fundamento de dicho método radica en tomar como estimadores de los parámetros desconocidos aquellos valores que hacen máxima la Verosimilitud de la muestra, o sea la probabilidad de haber observado precisamente los datos que se han obtenido.
 - Este estimador es asintóticamente insesgado y de varianza mínima
 - Esta maximización se realiza a través de métodos de optimización.

Odds

- La regresión logística modela la probabilidad de que la variable de respuesta pertenezca a una clase indirectamente mediante el logaritmo de odds.
- Un evento tiene probabilidad 0.8 de ocurrencia. La probabilidad de que esta no ocurra es 0.2. Un *odds ratio*, es el ratio entre la probabilidad de que sea de una clase versus que no lo sea.
 - $\text{Odds} = 0.8 / 0.2 = 4$
 - Se esperan 4 eventos verdaderos por cada evento falso.

IC parámetros

- Basado en la curva normal

$$(\hat{\beta}_1 - z_{1-\alpha/2}\sigma(\hat{\beta}_1), \hat{\beta}_1 + z_{1-\alpha/2}\sigma(\hat{\beta}_1))$$

donde $z_{1-\alpha/2}$ es el valor que deja una probabilidad $1 - \alpha/2$ a su izquierda en una $N(0, 1)$.

$$\left(e^{\hat{\beta}_1 - z_{1-\alpha/2}\sigma(\hat{\beta}_1)}, e^{\hat{\beta}_1 + z_{1-\alpha/2}\sigma(\hat{\beta}_1)} \right)$$

Test de Wald

- Permite determinar si cada coeficiente del modelo es significativo o no.

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned} \quad W = \frac{\hat{\beta}_1}{\sigma(\hat{\beta}_1)}$$

- Este estadístico sigue una distribución chi-cuadrado con 1 grado de libertad.
- Como en todo test de hipótesis se rechaza la hipótesis nula si el p-valor del estadístico es pequeño.

Bondad de ajuste

- Similar al coeficiente de determinación de la regresión lineal, es decir la varianza explicada por el modelo. Una vez que este sea significativo.
 - **-2 log de la verosimilitud:** mide hasta qué punto un modelo se ajusta bien a los datos. Cuanto más pequeño sea el valor, mejor será el modelo.
 - **R cuadrado de Cox y Snell:** Sus valores oscilan entre 0 y 1 (tiene un valor máximo inferior a 1, incluso para un modelo perfecto).
 - **R cuadrado de Nagelkerke:** es una versión corregida de la R cuadrado de Cox y Snell y cubre el rango completo de 0 a 1.
- La prueba de Hosmer-Lemeshow es otro método para estudiar la bondad de ajuste del modelo de regresión logística que consiste en comparar los valores previstos (esperados) por el modelo con los valores realmente observados.

Supuestos

- **No colinealidad o multicolinealidad:** Predictores independientes.
- **Relación lineal entre los predictores numéricos y el logaritmo de *odds* de la variable respuesta**
- **No autocorrelación (Independencia):** Las observaciones deben ser independientes entre sí.
- Otros puntos importantes son:
 - Valores atípicos
 - Tamaño de la muestra
 - Parsinomia: capacidad del modelo de explicar con mayor precisión la variabilidad observada mediante el menor número de features.

Feature selection

Se relaciona con overfitting, ya que permite reducir y no saturar los modelos y producir un efecto indeseado.

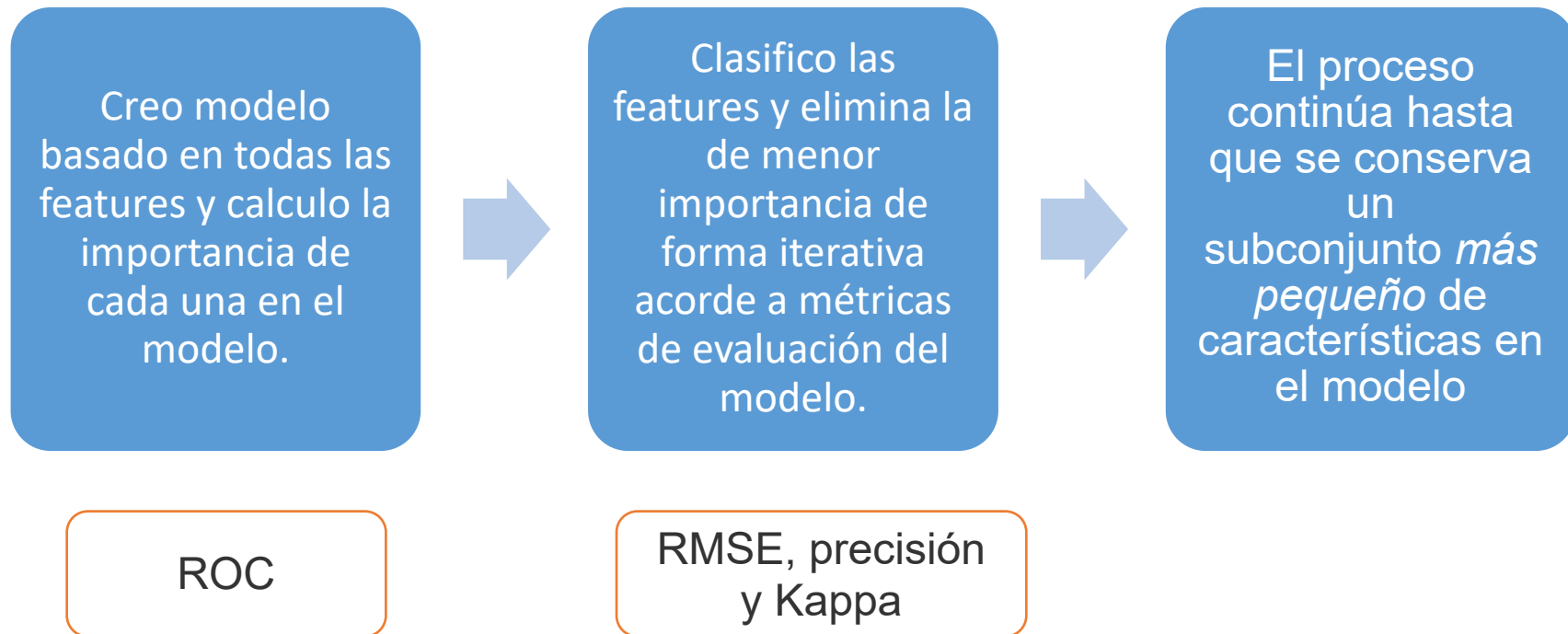
Para esto hay métodos de selección de variables.

Esto permite aumentar el rendimiento del modelo en set de datos multidimensionales.

Recursive feature elimination (RFE)

- Algoritmo ampliamente utilizado para seleccionar características que son más relevantes para predecir la variable objetivo en un modelo predictivo, ya sea regresión o clasificación.
- RFE aplica un proceso de selección hacia atrás para encontrar la combinación óptima de características.

Pasos RFE



Recursive feature elimination Cross Validation (RFECV)

- Funciona exactamente igual que el algoritmo anterior pero utilizando Cross Validation en el proceso.
- Utiliza la validación cruzada para conservar las mejores features de rendimiento.
- Estos procesos los utilizaremos no solo con regresión logística sino que en otros algoritmos de Machine Learning.

Regresión Logística Multinomial

- Extensión de regresión logística
- Versión modificada de la regresión logística que predice una probabilidad multinomial (es decir, más de dos clases) para cada ejemplo de entrada.
- Cambiar la regresión logística de probabilidad binomial a multinomial requiere un cambio en la función de probabilidad y un cambio en la salida de un valor de probabilidad único a una probabilidad para cada etiqueta de clase.

Prediciendo sobrevivientes del Titanic

