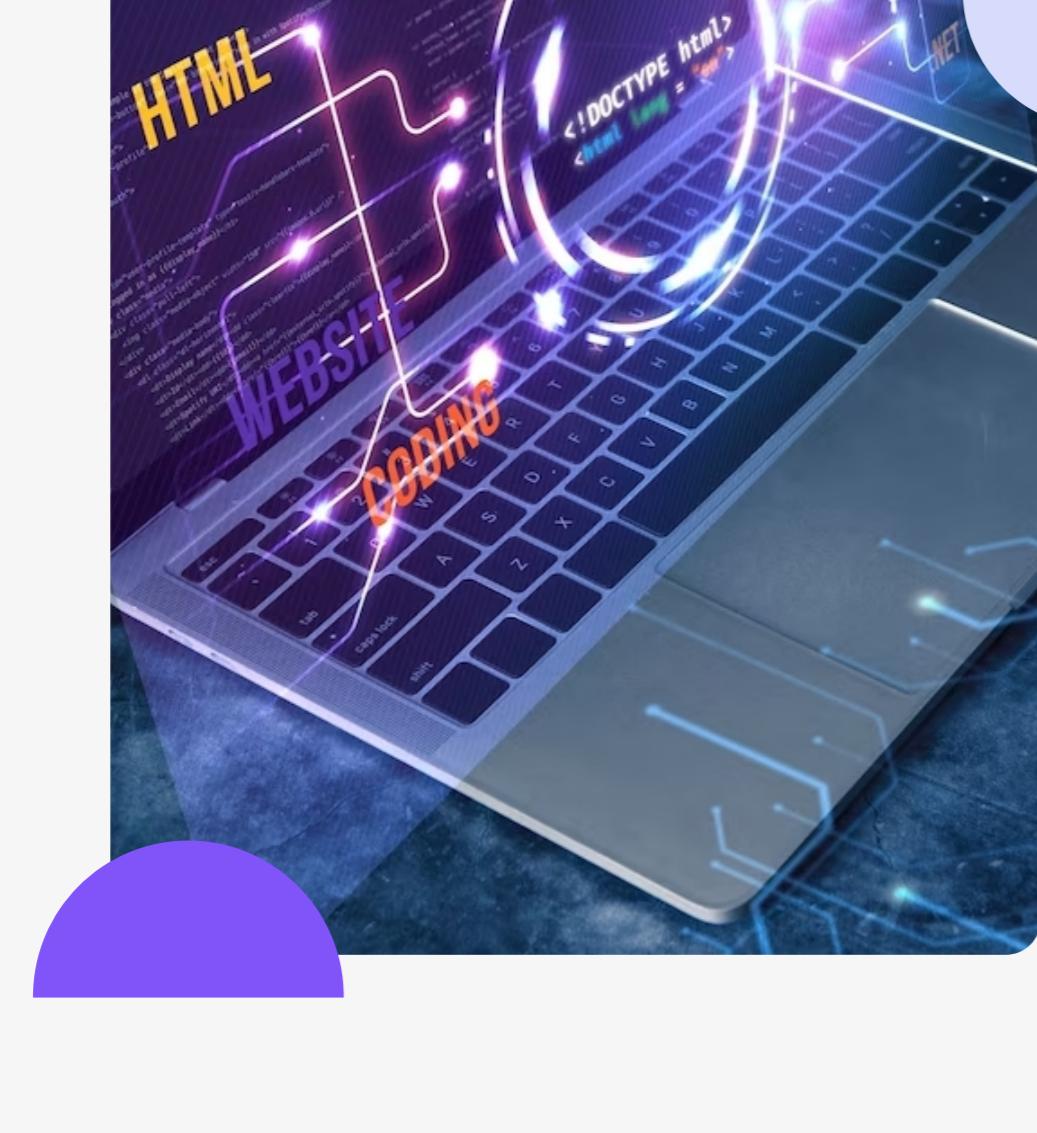


Algoritmo de regresión

Este algoritmo se basa en un método estadístico para la predicción de clases que sean binarias, el resultado o las variables objetivo deben ser de naturaleza **dicotómica**; la dicotomía hace referencia a que solo puede haber dos clases posibles. Dentro de los usos que se pueden encontrar en este tipo de herramienta, está la posibilidad de determinar enfermedades como el cáncer; en el cálculo de la probabilidad de una situación determinada, por ejemplo, para determinar si los clientes de determinada empresa comprarían un nuevo producto o se lo comprarían a la competencia.

La regresión logística es un algoritmo bastante simple y ampliamente utilizado para realizar la clasificación de dos clases; la implementación de este algoritmo es fácil y podría usarse como base principal para la solución de problemas en los que se busque clasificar el resultado de forma binaria. Básicamente, la regresión logística realiza la descripción y hace una estimación entre la relación que existe de una variable que es dependiente y las que son independientes.



La fórmula que se utiliza para la función logística se conoce también como la función sigmoide, aquí se establece relación entre la variable dependiente y las independientes.

$$f(x) = \frac{1}{1+e^{-x}}$$

x representa un número real. En esta ecuación, se puede observar que cuando x tiende al menos infinito, el cociente tiende a cero, y, por el contrario, cuando x tiende al infinito, el cociente dará como resultado tendencia a uno.

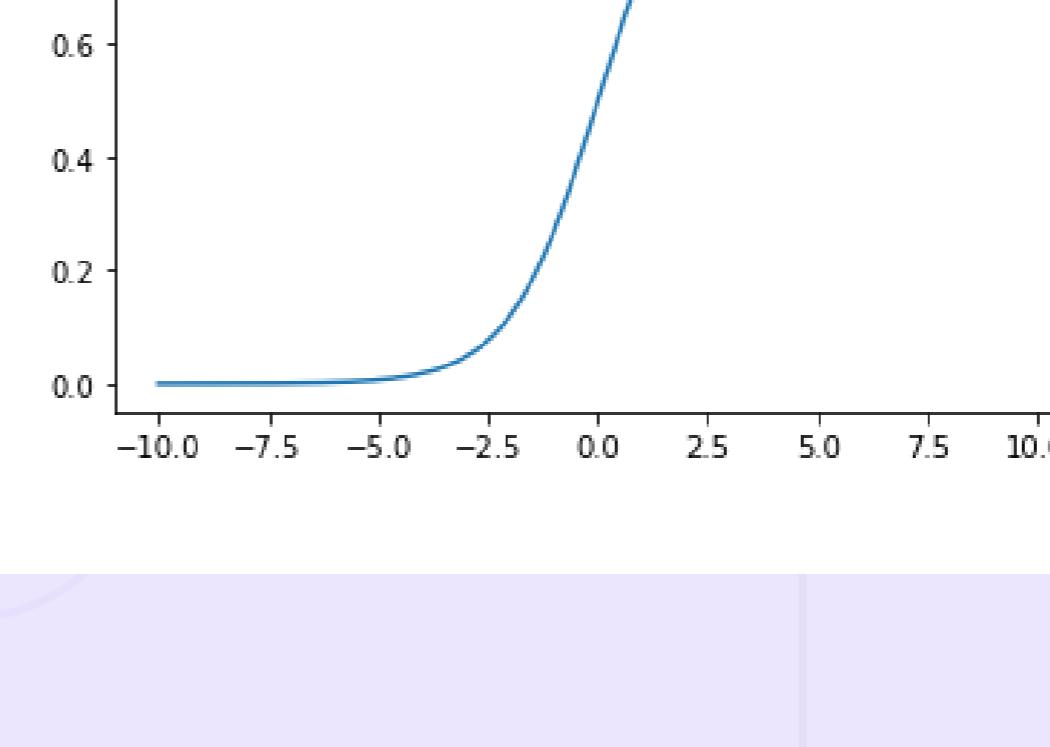
El resultado gráfico de esta fórmula se muestra en forma de S, en la cual los valores que toma son de 0 y 1, nunca se tomará un valor por fuera de estos rangos. A continuación, las siguientes líneas de Python ayudan a simular esta gráfica.

```
import numpy as np
import matplotlib.pyplot as plt

x = np.arange(-10, 10, 0.1)
y = 1/(1+np.exp(-x))
plt.plot(x, y)
plt.show()
```

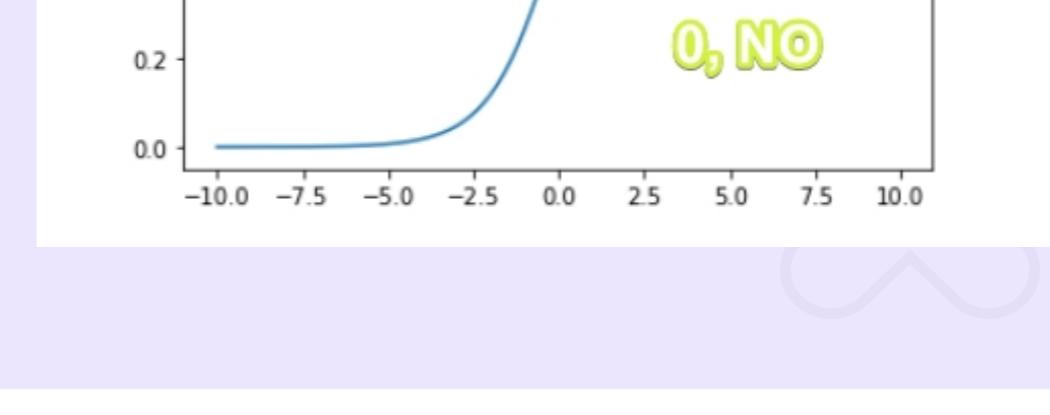
Con la ayuda de las librerías *numpy* y *matplotlib*, se realiza la simulación de la gráfica sigmoide, en la que se pasa a la variable x números desde -10 a 10, en intervalos de 0.1, y como se mencionó anteriormente, si este valor tiende al infinito negativo, el resultado tiende a 0, y si lo hace hacia el positivo, el resultado tiende a uno. (Véase Figura 1)

Figura 1 Gráfica de la función sigmoide

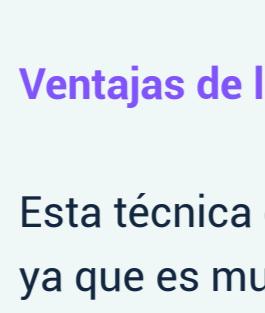


Para los resultados que se obtengan superiores a 0.5, se determina ese resultado como un 1 o respuesta afirmativa, mientras que si el resultado se encuentra por debajo de 0.5, se puede clasificar como 0 o respuesta negativa; asimismo, los resultados se pueden interpretar como el porcentaje de la probabilidad. Por ejemplo, en un problema de salud en el que se quiera determinar la enfermedad de cáncer en pacientes tratados y el resultado que se obtiene es de 0.85, se podría interpretar que ese paciente tendría un 85 por ciento de probabilidad de contraer esa enfermedad.

Figura 2 Resultados cero y uno

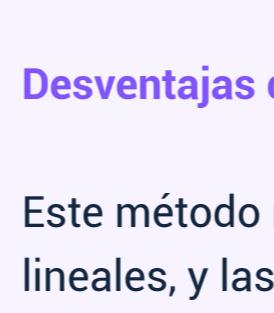


A continuación, se mencionan algunas ventajas y desventajas de esta técnica



Ventajas de la regresión logística

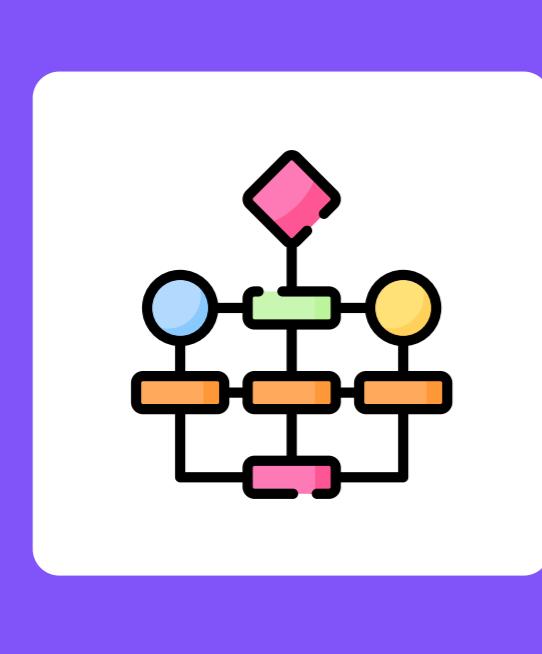
Esta técnica está siendo muy utilizada por los científicos de datos, ya que es muy eficiente y fácil de implementar, no se necesita disponer de grandes recursos a nivel computacional ni para la parte de entrenamiento y ejecución, los resultados son muy fáciles de interpretar, siendo quizás esta la más importante de sus ventajas.



Desventajas de la regresión logística

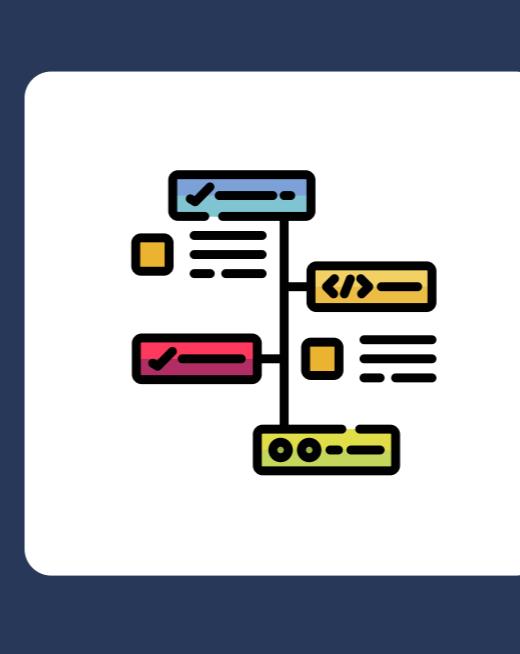
Este método no se puede utilizar en la solución de problemas no lineales, y lastimosamente existen en la actualidad muchos sistemas que son no lineales; de igual manera, solo se puede utilizar la regresión logística para dar solución o predecir problemas en los que su variable objetivo sea categórica, la regresión logística no es considerada como uno de los algoritmos más potentes que actualmente existen.

Por otro lado, se pueden encontrar otros tipos de regresión logística, entre los que están:



Multinomial

La variable objetivo tiene tres o más categorías nominales, no se especificaría ningún orden; por ejemplo, predecir si los clientes de un restaurante elegirán entre un tipo de vino específico o pedirán comida vegetariana, vegana o carnes.



Ordinal

Igualmente, puede haber más de tres variables como en la multinomial, pero estas siguen un orden; por ejemplo, la calificación de 1 a 5 del restaurante.



Se puede concluir que este algoritmo es muy fácil de implementar e interpretar, su coste a nivel computacional es muy bajo, pero también se puede decir que no funciona bien para problemas que no son linealmente separables.