One-Hot Encoding y Label Encoding

One-Hot Encoding

La técnica One-Hot Encoding es un método para convertir variables categóricas en una representación numérica que puede ser utilizada por algoritmos de aprendizaje automático. Se aplica cuando se tienen variables categóricas nominales, es decir, aquellas que no tienen un orden inherente (como colores, tipos de productos, géneros, etc.).

¿Cómo funciona?

1. Identificación de categorías: Se toman todas las categorías únicas de la variable.
2. Creación de nuevas columnas: Se genera una columna binaria (0 o 1) por cada categoría.
3. Codificación: Para cada observación, se coloca un 1 en la columna correspondiente a su categoría y 0 en las demás.

Ejemplo

Color	Rojo	Verde	Azul
Rojo	1	0	0
Verde	0	1	0
Azul	0	0	1
Rojo	1	0	0

¿Cuándo se aplica?

- Cuando se tienen variables categóricas nominales en algoritmos que solo trabajan con datos numéricos.
- En modelos como regresión lineal, redes neuronales y árboles de decisión, donde las variables categóricas no pueden ser interpretadas directamente.
- Cuando las categorías no tienen un orden lógico (si lo tuvieran, se podría usar Label Encoding en su lugar).

Consideraciones

- Puede generar alta dimensionalidad si hay muchas categorías únicas, lo que puede afectar el rendimiento del modelo.
- En algunos casos, se usa One-Hot Encoding con eliminación de una columna para evitar la multicolinealidad en modelos lineales.

Label Encoding

Label Encoding es una técnica utilizada en aprendizaje automático para convertir variables categóricas en valores numéricos. Se usa cuando una variable categórica tiene valores que pueden representarse mediante números enteros sin perder información significativa.

¿Cómo funciona?

Cada categoría en la variable categórica se asigna a un número único. Por ejemplo:

Color	Label Encoding
Rojo	0
Verde	1
Azul	2

¿Cuándo se aplica?

- Cuando hay una relación ordinal en los datos (es decir, un orden lógico entre las categorías), como en:
  - Tamaño de ropa: Pequeño (0), Mediano (1), Grande (2)
  - Nivel de educación: Primaria (0), Secundaria (1), Universidad (2)
- Cuando se quiere reducir la dimensionalidad en comparación con One-Hot Encoding, especialmente en variables con muchas categorías.

Desventajas

- Si se usa en variables no ordinales, los modelos pueden interpretar que hay una relación numérica entre las categorías cuando no la hay. Por ejemplo, en el caso de los colores, el modelo podría pensar que Azul (2) es mayor que Rojo (0), lo cual no tiene sentido.