18/12/2023

Unit 06 – Data cleaning

Los modelos tiene 2 problemas a la hora de aceptar datos:

* No acceptan datos
* No hacen datos categoricos (strings)

Datos nulos

Hay 3 principals opciones para lidiar con datos nulos

* Eliminar la columna entera. Interesante cuando hay muchos nulos en esa columna
* Eliminar la fila entera. Es una solución sencilla, pero si llega a producción un dato con nulo, el modelo no será capaz de funcionar para ese caso. Si se toma esta decisión, es muy importante indicarlo en el modelo
* Inferir el valor: con la media, mediana, mínimo…

Datos categóricos

La mayoría de modelos no aceptan datos categóricos o no es lo ideal. Para transformar datos categóricos en numéricos

LabelEncoder: Asignar un numero a cada categoría. No tiene orden.

OrdinalEncoder: se le puede agregar un orden predefinido cuando se quiere que el modelo entienda ese orden

OneHotEncoder: divide las categorías de una columna en una columna para cada categoría

Preparar datos

El columna transformer permite definir los transformadores que se quiere aplicar a las columas