Manejo de datos nulos o faltantes

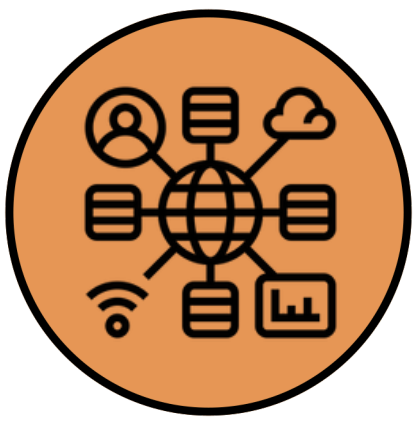
LECCIÓN 3

Tiempo de ejecución: 1 hora	
PLANTEAMIENTO DE LA SESIÓN	**MATERIALES**
En los conjuntos de datos suelen existir faltantes y nulos, es decir, valores que por algún motivo se perdieron o no fueron capturados. Según la situación se debe dar solución con diferentes técnicas, siempre dependiendo del contexto de los datos.	

El manejo de datos nulos o faltantes es una parte crucial del proceso de limpieza de datos en ciencia de datos. Los datos faltantes pueden surgir por diversas razones, como errores de entrada, fallos en la recopilación de datos o simplemente porque cierta información no está disponible.

Para manejar los datos nulos es necesario identificarlos junto con su contexto. Esto puede implicar visualizar los datos, calcular la cantidad de valores faltantes en cada variable o utilizar métodos como la función isnull() en bibliotecas como pandas en Python para identificar valores nulos en un conjunto de datos.

Si los datos faltantes son pocos y no afectan significativamente la integridad del conjunto de datos o el análisis posterior, a veces es aceptable simplemente eliminar las filas o columnas que contienen valores nulos. Sin embargo, esto debe hacerse con cuidado para no perder información importante.



Otra forma de lidiar con los datos faltantes y su contexto es mediante la imputación. Esto es, rellenar los valores faltantes con sustitutos plausibles, como por ejemplo colocar la media o la moda de la variable respectiva. Esto no afectará la distribución de probabilidad de la variable y permitirá conservar otras columnas con datos que pueden ser útiles.

Otra forma más sofisticada de realizar la imputación es por medio de modelos de regresión, como por ejemplo k-NN o árboles de decisión, para predecir los valores faltantes en función de otras variables.