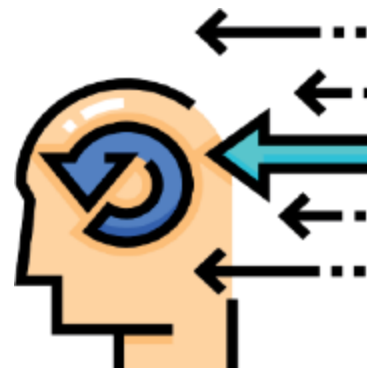




Módulo 5

Sesión N° 4



ACTIVIDAD:



Diagnóstico de calidad de datos usando Pandas

Objetivo: Aplicar los conocimientos adquiridos sobre las dimensiones y principios de calidad de datos (ISO 8000), junto con el uso práctico de la librería Pandas, para identificar problemas comunes y proponer mejoras basadas en métricas objetivas.



Contexto

Descarga el archivo llamado [clientes_muestra.csv](#), proveniente de una exportación de CRM. Este contiene registros con problemas comunes como campos incompletos, duplicados y formatos inválidos.

Se espera que los estudiantes:

- Analicen el dataset con Pandas.
- Evalúen la calidad en base a métricas cuantitativas.
- Construyan una tabla de diagnóstico similar a la siguiente:



Resumen de calidad de datos:

Campo	Compleitud (%)	Validez (%)	Duplicados
email	95%	65%	3
nombre	85%	100%	0
id_cliente	100%	100%	0

Tiempo estimado: 30 minutos

Formato de ejecución: grupal





Requerimientos:

1. Abran el archivo clientes_muestra.csv en su editor (VSCode, Jupyter Notebook, Google Colab, etc.).
2. Usen el entorno virtual ya configurado previamente con Pandas. Si no lo tienes activo, reactívalo con:

workon amb_mod5

3. Creen un script llamado diagnostico.py e incluyan las siguientes operaciones:
 - o Cargar el archivo con `pd.read_csv()`
 - o Calcular porcentaje de completitud `(.notnull().mean() * 100)`
 - o Validar correos con una expresión regular
 - o Detectar duplicados con `.duplicated().sum()`
4. Construyan una tabla como la tabla de referencia, indicando:
 - o Qué campo tiene más errores
 - o Qué dimensión de calidad está más comprometida
 - o Qué consecuencias podría tener en el negocio
5. Compartan sus hallazgos de forma breve con el resto del curso.

