Clase 1 – Tipos de Datos y Generalización

■ Tipos de Datos en Machine Learning

- Numéricos (ej: edad, salario, temperatura).
- Categóricos (ej: género, estado civil, color).
- Texto (ej: comentarios, artículos, reseñas).
- Imágenes (ej: dígitos escritos a mano, fotos).
- Audio y video (ej: grabaciones, cámaras de seguridad).
- Estructurados: tablas con filas/columnas.
- No estructurados: texto libre, imágenes, audio.

■ Relación Problema – Datos – Modelo

El tipo de dato condiciona el tipo de modelo a utilizar. Ejemplos:

- Datos numéricos → regresión lineal, árboles de decisión.
- Datos categóricos → clasificación (ej: Naive Bayes, Random Forest).
- Texto → modelos de NLP (ej: Bag of Words, Transformers).
- Imágenes → redes neuronales convolucionales (CNN).

■ Generalización

La generalización es la capacidad de un modelo para aprender patrones que funcionen en datos nuevos, y no solo en los datos de entrenamiento. Un buen modelo generaliza correctamente.

■■ Overfitting vs Underfitting

- Overfitting: el modelo memoriza demasiado los datos de entrenamiento, pierde capacidad de predecir datos nuevos.
- Underfitting: el modelo es demasiado simple y no logra aprender los patrones importantes.
- Objetivo: encontrar un balance adecuado entre ambos.

■ Actividad en Clase

1. Revisar un dataset real (ejemplo Titanic).
2. Identificar qué tipos de datos contiene.
3. Discutir en grupos: ¿qué pasaría si un modelo memoriza en lugar de aprender?