Missing Indicator & Random Sample Imputation

➢ Handling Missing Data:

This part of the lecture introduces **two key techniques** for handling missing values in datasets:

---

1. Random Sample Imputation:

- ***Concept:*** Replaces missing values with **random values** taken from the **same variable's non-missing values**.
- ***Why use it?*** It maintains the **original distribution** of the data, unlike mean/median imputation which can distort variance.
- ***Risk:*** It can still introduce randomness and **overfit** if the sample size is small.

---

2. Missing Indicator Method:

- ***Concept:*** Adds a new **binary column (0/1)** that indicates whether the value in the original column was missing.
- ***Purpose:*** Helps models **capture the "missingness" pattern** — which might carry predictive information.
- Often **used alongside imputation** (like mean or median) for better model performance.

---

3. Auto Value Selection for Imputation:

- **Tools** like SimpleImputer in **scikit-learn** can **automatically detect and apply** an imputation strategy (e.g., mean, median, most frequent).
- This is helpful for **scaling to large datasets** or pipelines.

💡 When to Use These Techniques?

- Use **random imputation** to preserve variance when you don't want to distort distribution.
- Use the **missing indicator** when **missingness itself** might be informative (e.g., missing age on Titanic may correlate with survival).
- Combine both when building models that can benefit from this extra information.
