Data leakage

What is Data Leakage?

Data Leakage occurs when **your model accidentally gains access to information during training that it wouldn't have at prediction time.**
This results in **unrealistically high performance** during training or testing — but **fails in real-world predictions**.

Why It's Dangerous:

- **Makes your model look perfect** in testing.
- **Fails miserably** in production (real unseen data).
- **Invalidates your evaluation metrics.**

✅ **Real-World Examples:**

✅ **Target leakage:**

Using a feature that contains direct information about the target.

Example:
You are predicting whether a patient has diabetes.
But one feature is: glucose_level_after_diagnosis → this happens **after** the diagnosis.

Fix: Only use features that are **available before the outcome** is known.

✅ **Train-Test contamination:**

Information from the **test set accidentally leaks** into the training set.

Example:
You do feature scaling (e.g., StandardScaler) **before** splitting data into train/test.
The mean/std is computed using **all data**, including test!

Fix: Always split the dataset first, then apply transformations **separately** on train and test.

✅ **Using future data in time series**

Example:
In time series forecasting, using future prices to predict current values.

Fix: Always ensure **no data from the future leaks into the past**.

Type of Leakage	**Description**	**Solution**
Target Leakage	Feature uses future/target info	Drop such features
Preprocessing Leakage	Fit transformation before split	Split first, then transform
Temporal Leakage	Using future time data	Use only past data for training

Golden Rule:

"Never use any information during training that would not be available when making predictions.

Real-world Data vs Test Data

- **Test data:** Validate the model by dividing it into 10,000 previous patient reports
- **Real-world:** A new patient's report came from a completely different hospital

If the format changes in the new report, or there are new types of symptoms — the model may become confused.