Repeated observations, group leakage and predictive baselines — review sheet

Use this after attempting the worked examples and practice. Explain each claim in words before calculating.

Key ideas

- Define the target before choosing the split.
- Audit both rows and groups; exact-row separation is insufficient for unseen people.
- All fitted preprocessing belongs inside the training boundary.
- Baselines and class-specific metrics make the claim inspectable.

Common errors and repairs

Treat a perfect contaminated score as new-person performance. It measures a task with already represented groups; use the target-matched protocol.

Fit the scaler on the full table. Fit on training rows and reuse the stored transform.

Treat class balance in this generator as population prevalence. It is a teaching design choice.

Retrieve without your notes

- What is the proper split's shared-group count?
- Why does matching training row counts help?
- Does grouping by person answer every site-shift question?

Readiness check

- Define a prediction target and its appropriate unit of separation.
- Audit both row-index and group-identity overlap in a proposed split.
- Fit transparent baselines with training-only preprocessing and interpret class-specific metrics.
- Explain why a strong contaminated score does not estimate performance on new groups.