# Regression, classification, and multivariate data — review sheet

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

## Key ideas

- Prediction time determines legitimate features.
- Fit transformations on training rows and reuse them on test rows.
- A baseline gives a metric a useful comparison.
- Precision and recall have different denominators.
- PCA preserves input variance; it does not choose labels.
- Cluster numbers are arbitrary and need scientific interpretation.

## Common errors and repairs

**Use the measured target as a predictor of its own label.** Keep only features available at the intended prediction time.

**Fit preprocessing on all rows.** Fit on training rows and apply the stored transform to held-out rows.

**Interpret a cluster number as a known class.** Compare partitions and labels only with an explicit evaluation question.

**Tune after seeing the test score and keep calling it untouched evaluation.** Use training-only validation for choices and a separate final evaluation.

## Retrieve without your notes

- What are the units of RMSE for a yield measured in mg?
- What denominator defines precision?
- Does a sign flip of PC1 change the projection axis?

## Readiness check

- Declare features, a continuous target and a binary label without leaking the outcome into inputs.
- Fit preprocessing and models on training rows, then report held-out comparisons to constant baselines.
- Calculate RMSE, accuracy, precision and recall with the correct units and denominators.
- Interpret PCA as a variance-based projection and two-center clustering as a distance-based grouping.
- Distinguish supervised prediction from unsupervised structure and state limits of a synthetic benchmark.