A4 MACHINE LEARNING · A4.2 · HL ONLY

Data preprocessing

Turning messy data into model-ready input: **data cleaning**, **feature selection**, and **dimensionality reduction**. HL only.

01 Why preprocess

Principle	Garbage in, garbage out.
Goal	Clean, consistent, model-ready data.
Pipeline	Clean → transform → select → reduce.
Impact	Often beats a fancier algorithm.

02 Data cleaning

Missing	Impute (mean/placeholder) or remove.
Outliers	Trim (remove) or cap (clamp).
Duplicates	Remove repeated records.
Noise	Fix bad formats and invalid values.
Spot	Use box plots to find outliers.

03 Cutting features · selection vs reduction

- **Feature selection**

  Keep a relevant **subset of the original** features and drop the rest. Stays easy to interpret.

- **Dimensionality reduction**

  Combine features into fewer new components (such as PCA). Powerful, but harder to interpret.

04 Transformation

Normalise	Scale values to a common range.
Encode	Turn categories into numbers.
Why	Many algorithms need numeric, scaled input.
Curse	Too many features slow training, cause overfit.

05 Worked example: clean a row

Missing email	Use a placeholder, not row deletion.
Negative income	An error: correct or remove it.
Duplicate 002	Delete the repeated record.
Result	Consistent, trustworthy data.

06 Know the difference

Selection vs reduction	Keep a subset of original features versus combine them into fewer new ones.	FEATURES
Impute vs remove	Fill a missing value with a substitute versus dropping the record entirely.	MISSING
Trim vs cap	Remove an outlier versus clamping it to a sensible threshold.	OUTLIERS
Cleaning vs transformation	Fixing errors and gaps versus reshaping valid data into a model-ready numeric form.	STAGE

FINAL PASS BEFORE THE EXAM

Rapid exam tips

Eight things that lose marks in Paper 1 if you slip on them. A4.2 is HL only. Skim before you walk in.

01

Garbage in, garbage out: clean data beats a fancier model.

02

Missing values: **impute** (mean/median/placeholder) or remove.

03

Outliers: **trim** or **cap**. Don't delete genuine extreme values.

04

Feature selection keeps a subset of the original features.

05

Dimensionality reduction (PCA) combines features into fewer new ones.

06

Selection stays **interpretable**; reduction is harder to interpret.

07

Transformation: normalise/scale values and encode categories as numbers.

08

This whole subtopic is **HL only**.