Objective

We model the probability that an adverse reaction belongs to a given MedDRA System Organ Class (SOC) using patient, product, dose and administration information from the Canada Vigilance dataset. Let $\mathbf{x}$ be the vector of predictors and Y the SOC label:

$$\hat{y} = \arg\max_{\text{soc}} \Pr\big(Y = \text{soc} \mid \mathbf{x}\big), \qquad Y \equiv \text{SOC\_NAME\_ENG}.$$

Predictors (Inputs)

Demographics and anthropometrics

- `GENDER_CODE`: Encoded sex of the patient.
- `age_y`: Age in years.
- `weight_kg`: Body weight in kilograms.
- `height_cm`: Height in centimetres.

Exposure and dosing

- `routeadmin_eng`: Route of administration (e.g. oral, IV).
- `unit_dose_qty`: Numeric dose per administration.
- `dose_unit_eng`: Dose unit in English (e.g. mg, mg/kg, mL).
- `hours_between_medicament`: Time in hours between administrations.

Product and indication

- `indication_name_eng`: Therapeutic indication in English.
- `active_ingredient_name`: Active ingredient name(s) of the product.

Target (Output)

- `SOC_NAME_ENG`: MedDRA System Organ Class in English (e.g. “Cardiac disorders”, “Skin and subcutaneous tissue disorders”).

Modeling Note

This is a supervised multi-class classification problem with potential class imbalance across SOCs. Evaluation should therefore include macro-averaged metrics (e.g. macro-F1, macro-recall) and confusion matrices by SOC.