PYTHON CHO KHOA HỌC DỮ LIỆU

Giảng viên: Hà Minh Tuấn

THAM SỐ MẶC ĐỊNH CỦA DataPreprocessor

Nhóm thực hiện:

Đặng Đỉnh Đoàn 23280046
Phạm Thanh Uy 23280097

TP. Hồ Chí Minh, tháng 12 năm 2025

Mục lục

1 Giới thiệu

Dây là các tham số mặc định của class Preprocessor, khi User không truyền vào, sẽ tự động lấy các tham số này.

2 Nhóm đường dẫn & target – PreprocessorConfig

2.1 Đường dẫn file input/output

- **raw_data_path – đường dẫn file input thô:**
  - Nếu `None` → mặc định lấy `data/raw/diabetes.csv`
  - Có thể đổi thành:
    - `"data/raw/pima.csv"`
    - `"data/raw/my`$_d$*ataset.csv*”
    - `Hoặc đường dẫn tuyệt đối: r"E:\data\pima.csv"`
- **processed_train_path – đường dẫn file train đã xử lý:**
  - Nếu `None` → không tự lưu file processed
  - Nếu set:
    - `"pima_train_processed.parquet"`
    - `"train.parquet"` hoặc `"train.csv"`
- **processed_test_path – đường dẫn file test đã xử lý:**
  - Nếu `None` → không tự lưu file processed
  - Nếu set:
    - `"pima_test_processed.parquet"`
    - `"test.parquet"` hoặc `"test.csv"`
- **save_scaler_path trong scaler – nơi lưu scaler:**
  - `None` → không lưu scaler
  - Ví dụ: `"scaler.joblib"`, `"artifacts/scaler_pima.joblib"`, ...

2.2 Cột target

- **target_col – cột nhãn (label) của dataset:**
  - Hiện dùng: `"Outcome"`
  - Có thể đổi khi dùng dataset khác: `"Target"`, `"HasDiabetes"`, `"Label"`, ...

3 Hidden missing & Missing – hidden_missing_cols, missing

3.1 Hidden missing

- **hidden_missing_cols – các cột coi là missing ẩn:**
  - Hiện dùng: `["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]`
  - Có thể thêm/bớt cột:
    - Thêm `"Age"` nếu muốn xem Age=0 là missing
    - Bỏ `"SkinThickness"` nếu dataset khác không có
    - Đổi tên theo dataset mới, ví dụ: `"Glucose"` → `"GlucoseLevel"`

3.2 Missing numeric/categorical – missing

- **missing = {**
  - `"numeric_strategy": "median_by_outcome",`
  - `"categorical_strategy": "most_frequent",`

  }
- **numeric_strategy – các lựa chọn:**
  - `"median_by_outcome"`: Median theo từng nhóm Outcome 0/1, fallback sang median overall (hiện đang dùng)
  - `"median_overall"`: Median trên toàn bộ tập Train, không phân nhóm
- **categorical_strategy – các lựa chọn:**
  - `"most_frequent"`: dùng mode (hiện đang dùng)
  - `"constant"`: luôn điền "missing"hoặc 1 giá trị cố định

4 Outlier & Scaler – outlier, scaler

4.1 Outlier – outlier

- **outlier = {**
  - `"numeric_cols": None,`
  - `"method": "iqr",`
  - `"strategy": "winsorize",`
  - `"iqr_factor": 1.5,`

  }
- **numeric_cols – cột xử lý outlier:**

  - `None` (hiện dùng): tự lấy tất cả numeric trừ target_col
  - Hoặc list rõ ràng: `["Glucose", "Insulin", "BMI"]`
- **method – cách phát hiện outlier:**
  - `"iqr"` (hiện dùng): dùng IQR (Q1 ± factor × IQR)
  - Có thể thêm: `"zscore"`, `"none"`
- **strategy – cách xử lý outlier:**
  - `"winsorize"` (hiện dùng): kẹp giá trị về [lower, upper]
  - `"flag"`: thêm cột is_outlier_<col> 0/1, không thay đổi giá trị
  - `"none"`: bỏ qua detect_outliers
- **iqr_factor – hệ số IQR:**
  - Mặc định: 1.5
  - Có thể giảm (1.0) → nhiều điểm bị coi là outlier hơn
  - Có thể tăng (2.0, 3.0) → ít điểm bị coi là outlier hơn

4.2 Scaler – scaler

- **scaler = {**
  - `"type": "standard",`
  - `"exclude_cols": [],`
  - `"save_scaler_path": None,`

  }
- **type – loại scaler:**
  - `"standard"` (hiện dùng): StandardScaler
  - `"minmax"`: MinMaxScaler (0–1)
  - `"none"`: bỏ qua scale
- **exclude_cols – cột không scale:**
  - Ví dụ: `["Outcome"]`, hoặc `["Outcome", "Pregnancy_high"]`
  - scale_features vẫn tự exclude target_col
- **save_scaler_path – đường dẫn lưu scaler:**
  - None → không dump scaler
  - Ví dụ: `"scaler.joblib"`, `"artifacts/pima_scaler.joblib"`

5 Encoding – encoding

- **encoding = {**
  - `"strategy": "onehot",`
  - `"handle_unknown": "ignore",`

  }
- **strategy – cách encode categorical:**
  - `"onehot"` (hiện dùng): pd.get_dummies, align cột Train/Test
  - `"label"`: LabelEncoder từng cột
  - `"none"`: bỏ encode (data đã numeric)
- **handle_unknown – xử lý giá trị lạ:**
  - `"ignore"` (hiện dùng): align cột, thêm 0 nếu thiếu
  - Có thể mở rộng: `"error"` → raise exception

6 Feature engineering – feature_engineering

- **feature_engineering = {**
  - `"enable": True,`
  - `"create_bmi_category": True,`
  - `"create_age_group": True,`
  - `"create_pregnancy_flag": True,`
  - `"create_interactions": True,`
  - `"bmi_col": "BMI",`
  - `"age_col": "Age",`
  - `"pregnancies_col": "Pregnancies",`
  - `"glucose_col": "Glucose",`
  - `"insulin_col": "Insulin",`

  }
- **enable – bật/tắt toàn bộ feature engineering:** True/False
- **create_... – bật/tắt từng feature:** True/False
- **..._col – đổi tên cột nguồn nếu dataset khác:** ví dụ `"BMI"` → `"BodyMassIndex"`, ...

7 Split train/test – split

- **split = {**
  - `"test_size": 0.2,`
  - `"random_state": 42,`
  - `"stratify": True,`

  }
- **test_size – tỉ lệ test:** 0.2 (hiện dùng), có thể 0.1, 0.25, 0.3 ...
- **random_state – seed shuffle:** bất kỳ int, cố định để reproducible
- **stratify – stratify theo nhãn y:** True (giữ tỉ lệ 0/1), False (ngẫu nhiên)

8 Các "tùy chọn mềm" qua hàm/method

- **load_data(path=None):** truyền trực tiếp đường dẫn file, bỏ qua config.raw_data_path
- **run_full_preprocessing(df=None):** nếu df=None → load data theo config, nếu df có sẵn → dùng trực tiếp
- **Chọn class preprocessor:** DataPreprocessor(config=...) → bản gốc, LoggingPreprocessor(config → có log missing summary