Avir MLE Time Series Prediction Task

Predict Next-Week Polypropylene Price Direction

Objective

Predict whether next week's **polypropylene (PP)** price will move **Up or Down** relative to the current week.

Data

- **Core series:** Weekly PP prices (2021–2025, provided).
- **Optional related inputs** (weekly or resampled daily → weekly):
  - **Crude oil** (Brent/WTI)
  - **Naphtha**
  - **Natural gas**
  - **Foreign exchange rates (USD-based)**
- **Public data sources allowed:** e.g., Investing.com, EIA, FRED, exchange websites (cite all sources used).

Target Definition

- **Binary label:**
  (y_t = 1) if (PP_{t+1} > PP_t); else (y_t = 0).
- **Optional neutrality band:** Define a no-change zone (e.g., |ΔPP| < 0.5%) → label as neutral (can be dropped or merged with Down).

Feature Engineering

Keep the feature set **lean but interpretable**:

- **Technical indicators:** RSI, SMA/EMA, MACD, rolling volatility, momentum, weekly returns.
- **Price spreads/ratios:** PP–naphtha, Brent–WTI, propylene–naphtha.
- **Lags & seasonal patterns:** Include 1–m week lags (no look-ahead).
- **Optional:** Include exogenous variables (e.g., crude oil) lagged or differenced.

Ensure all features are computed within each training fold to prevent leakage.

Modeling Methods

- Any **ML/DL approach**.
- Ensembling is allowed.
- Document:
    - Total feature count
    - Key hyperparameters

Validation & Holdout

- **Time-aware validation:** Expanding or blocked time-series split.
- **Holdout period:** Reserve the latest segment (e.g., last 6 months) for final testing only.
- Clearly report training/validation/test boundaries.

Metrics

Minimum required:

- **Accuracy**
- **F1 score**
- **Confusion Matrix**

Deliverables

1. **Reproducible codebase in Github:**
    - Main training script
    - Environment or requirements.txt
2. **Short report (≤2 pages):**
    - Data sources & preprocessing
    - Feature design & rationale
    - Validation setup
    - Key results (metrics + interpretation)
    - Limitations & next steps
3. **Inference script:**
    - Inputs: latest data
    - Outputs:
        - P(Up)
        - Predicted class (Up/Down/Neutral)
        - Top feature drivers or importances

Acceptance Criteria

- Code quality
- Model **outperforms a naïve baseline** (e.g., last-sign persistence) on the holdout.
- The validation procedure is **transparent, time-consistent, and leak-free**.